MEASUREMENT IN HIGHER EDUCATION

WOOD

The Library of
St. Olaf College
Northfield, Minn.
Accession No. 21329
Class 371.2 Vol.

DEC 12 1923
371.2

2 20

MEASUREMENT AND ADJUSTMENT SERIES

EDITED BY LEWIS M. TERMAN

MEASUREMENT IN HIGHER EDUCATION

BY BEN D. WOOD

Assistant Professor of Collegiate Educational Research in Columbia College: Assistant to the Dean of Columbia College

Yonkers-on-Hudson, New York

WORLD BOOK COMPANY

1923

BF431
W64
1923

WORLD BOOK COMPANY

THE HOUSE OF APPLIED KNOWLEDGE

Established 1905 by Caspar W. Hodgson
YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Scientific method in education involves the careful measurement of each child's ability to learn and of the amount that he has learned. It also involves adjustment of organization, subject matter, and methods of instruction to the varying needs and abilities of pupils. This book is the first of a series that will set forth the value, technique, and applications of educational measurement and adjustment. It will present this subject from the point of view of higher education

MAS : WMHE–1

Copyright 1923 by World Book Company
Copyright in Great Britain
All rights reserved
PRINTED IN U.S.A.

TO

D. G. W. AND A. W. W.

EDITOR'S NOTE

At the request of the author the usual brief introductory statement by the editor has in this volume been replaced by a more extended discussion which appears as Chapter I. In acceding to the author's wishes in this regard the editor has been influenced by the conviction that the significance of the contribution which this book makes to higher education is so great as to justify its introduction to the American Educational Public in a more attention-compelling manner than is possible in the traditional type of editorial statement.

Lewis M. Terman.

ACKNOWLEDGMENTS

The writer's unmeasured thanks are due, in the first place, to Professor E. L. Thorndike of Teachers College, whose generous counsel has been mainly responsible for whatever scientific merit this volume may possess.

In the second place, acknowledgment is due to Columbia College. This book would have been quite impossible without the active coöperation of the administration of Columbia College in providing an ideal laboratory and in appropriating generous funds for the prosecution of the research on which it is based.

Dean Herbert E. Hawkes of Columbia College, Professor J. J. Coss, Executive Officer of the Staff in Contemporary Civilization, and Professor A. L. Jones, Director of Columbia University Admissions, have had so large and indispensable a part in the production of the book that it is only by their magnanimity that it is issued under the sole authorship of the writer.

The writer wishes to express his indebtedness to many others, too numerous to mention specifically, who have helped in carrying out this research; but special thanks are due to the Staff in Contemporary Civilization, the members of which are largely responsible for the experiment with the new type content examination described in Chapter VIII: Professors Carman, Coss, Evans, Kendrick, Moon, and Weld, Drs. Edman, Graper, Munro, Schneider, and Messrs. Burtt, Dewey, Friess, Gazley, Leigh, Randall, and Tugwell. Thanks are also due to Registrar Grant, to Assistant Registrar Fox, and to Assistant Director of University Admissions Chadwick. To Professor T. L. Kelley, formerly of Teachers College, now

of Stanford University, the writer is grateful for much helpful advice and inspiration given so disinterestedly during the last half-dozen years. Acknowledgment is made of the careful work of Miss G. A. Taylor and of Mrs. Anna W. Scully on the statistical tables and on the manuscript.

It is a matter not only of pleasure but also of pride to the author to make grateful acknowledgment to Raymond M. Weaver, Assistant Professor of English in Columbia University, for his friendship and encouragement during the writing of this book, and for his careful reading of the proofs in their entirety.

Finally, the writer wishes to express his deep appreciation of Professor Terman's generosity in consenting to write an introductory chapter for the book, in addition to the usual Editor's Introduction.

B. D. W.

Columbia University

CONTENTS

CHAPTER PAGE

I. INTRODUCTION (*Dr. L. M. Terman*) 1

II. THE COLUMBIA COLLEGE EXPERIMENT 12

Experiences of Columbia College with the intelligence test as an example of the application of principles equally valid for all colleges and high schools—Official status of intelligence tests in Columbia College—Use of the intelligence test as a criterion for admission to college

III. DESCRIPTION OF THE THORNDIKE EXAMINATION . . 22

Origin of the intelligence examination—Characteristics of the Thorndike Examination—Types of distributions for freshmen in Columbia College, juniors in Columbia College, and undergraduates in Stanford University—Intelligence standards in other colleges—Stability of mean and of S.D. of the Thorndike Examination—Reliability of Thorndike Examination—Indirect measures of reliability—Conduct of the examination—The prevention of cheating—Cost of the intelligence examination—Methods of keeping records

IV. VALIDITY OF THE THORNDIKE EXAMINATION . . . 63

Prediction of success and failure in college—Correlation of intelligence scores with college success—The scholarship score—Comparison of predictive values of new and old plans of admission—A combination criterion for admission—Other comparisons—Conclusion

V. ANALYSIS OF THE THORNDIKE EXAMINATION . . . 92

Analysis handicapped by variable time limits of Thorndike tests—Possible uses of analysis—Distributions of scores on eight Thorndike tests showing undistributed scores—Correlations between groups of Thorndike tests and various college courses—Reliabilities of grades in various college courses, and their influence on the correlations with intelligence scores—Comparative predictive values—Data on adjustment of academic load by the use of the intelligence test

VI. COLLEGE SUCCESS 107

What is college success—Characteristics of college marks—The basis of college marks—The means of deriving

CHAPTER PAGE

college marks—Personal estimate of the instructor—Written examinations—The purpose of grading students in college—Experimental evidence of the inaccuracy of school marks—The College Entrance Board examinations in algebra and geometry for June, 1921—The reliability of grades in Columbia College—Data on quality, quantity, and combined quality-quantity indexes of college success—Who can succeed in college—Conclusion

VII. Some Principles of Measurement 141

Thorndike's analysis of special difficulties of mental and social measurement—Objectivity—Reference to a defined zero point in terms of a defined unit—Definition of the function to be measured—Consistency—Variability or inconstancy of mental functions in the same individual—Comparability—Distinction between "power" and "achievement"—The principle of "equal exposure" in power or intelligence tests—Validity of indirect measurements—Scope of a valid test—Scope of a valid examination—Measurement distinguished from pedagogy—Summary

VIII. Experiment with a New Method of Content Examining 177

Plan and form of New Examination—Sample questions from New Examination in C.C.—Sample of old-type essay questions—Making and administration of New Examination—Statistical results—Types of distributions—Correlation results—Influence of personal acquaintance—Reliability of new and old methods—General considerations—Pedagogical value of new and old methods—Economy of the new method—Conclusions—Instructors' opinions on new method—The New Examination method adopted for C.C.—Effect of New Examination method on the reliability and significance of the index of college success—Extension of use of new method—Economics affected by new method

IX. The New Type Examination in Physics . . . 215

Adaptability of new type questions to physics—Samples of five new type physics quizzes—New type final semester examination in physics—Correlation results—Pedagogical applications of New Examination

X. The New Type Examination in Government . . 226

Administrative and pedagogical applications of objective and comprehensive examinations—Sample of new type examination in government—Reliability and significance of New Examination

CHAPTER PAGE

XI. The New Type Examination in Zoölogy . . . 240
Distributions of grades for two classes—Sample of new type examination in zoölogy—Special features of the zoölogy experiment—Reliability of New Examination in zoölogy

XII. The New Type Examination in Other Subjects . 250
Economics—Philosophy—Greek Art—History—English—Civil Engineering

XIII. Summary 274

Appendixes:

I. The Thorndike Examination 293
Sample questions of the Thorndike Examination—Instructions for giving the Thorndike Examination—Instructions for scoring—Range and significance of scores

II. Data on the Separate Parts of the Thorndike Examination 309
Distributions of scores of Parts I, II, and III—Distributions of scores of various groups of tests in the Thorndike Examination—Stability of the means and S.D.'s of the parts of the Thorndike Examination—Weighting of the parts of the Thorndike Examination

III. Suggested Improvements in the Thorndike Examination 320
Time limits—Undistributed scores—Objectivity—Amplification—Elimination—Simplification—Systematization—Chart of suggested revision

Bibliography 331

Index 335

21329

MEASUREMENT IN HIGHER EDUCATION

CHAPTER ONE

Introduction

By the Editor, Lewis M. Terman

Mental and educational tests have come into very general use in elementary and secondary schools throughout the country. There are a great many standard tests of various types intended for use in Grades 1 to 12, and the number of pupils given one or more of them in a year must run into the millions. After a considerable period of question and doubt, it is now everywhere admitted that the results of standard tests in our lower schools have proved their value in the classification of pupils according to ability, in educational guidance, and in measuring the results of our educational efforts.

Only recently, however, have the methods of tests and measurements been adopted for practical use in the institutions of higher learning, the colleges and universities, which were responsible in the first place for launching them upon the educational world. It must be admitted that American education is most conservative at the top, at least in respect to attitude toward innovations based upon psychological or pedagogical experiment. There are doubtless good and sufficient reasons why this is so. The average college teacher has had little professional training along psychological lines, and he is extremely preoccupied with his special subject of instruction. He is likely to feel a just and conscious pride in his immunity to the ephemeral fads which sweep periodically

over the educational horizon and leave so little but disappointment behind them. At the same time, once an educational reform has been fully demonstrated to rest upon a scientific basis there is no need to fear as to its ultimate reception in the average university. There may be a period of watchful waiting, but when the innovation has been accepted its influence is likely to be thoroughgoing and widespread.

Experiments with psychological tests of college students have been reported from time to time since the early work of Cattell back in the "nineties." For a long time, however, the work remained purely in the experimental stage. It is largely due to psychology's dramatic success in the Great War that we had such rapid progress in this field in recent years. Since 1918 scores of minor and several very important researches along this line have been published. Of such studies by far the most thoroughgoing and conclusive is that described in this volume.

The value and interest of this account, which the author has so ably and convincingly presented, go far beyond the facts and circumstances of the particular research concerned. Instead of a dry description of an experiment the author has given us a vital treatment of the entire measurement problem as it presents itself to our institutions of higher learning. His book is therefore an ideal introduction to the subject and ought to be read by every teacher in every university, college and normal school in the United States. Indeed, the issues which the author treats are so fundamental, and his methods of dealing with them have such wide applicability, that his book should prove no less interesting or useful to high school than to college teachers. In no essential respect do the problems of mental and educa-

tional measurement in the high school differ from those in the college. In both cases the canons of scientific procedure are exactly the same; the tactical difficulties to be overcome in instituting a measurement program are the same; the practical uses to be made of results are the same. There is hardly a fact or conclusion in the entire volume that is not as germane to the high school as to the college situation.

The editor believes that the publication of this volume will exert a profound influence upon higher education in the United States, an influence which may prove as real as that of Flexner's report on the more limited field of medical education. The problems with which he deals are problems of fundamental importance in the conduct of any institution of higher learning. His data have been painstakingly gathered; they have been treated by the most approved scientific methods, and the conclusions rest upon a solid foundation of fact.

That the author's data are based upon the *Thorndike* intelligence test and upon certain *Columbia* tests of academic achievement is of course irrelevant to the main purpose of the book. Far from setting up the tests which he used as being in any sense ideal models, the author is aware of their imperfections and has shown the way to improvements. The point is that even when the tests herein described have all been superseded by better ones, as in time they of course will be, the value of this book will not be appreciably impaired.

The reader untrained in statistical method may at first be somewhat mystified by the frequent use of mathematical terms and by the occasional reference to statistical formulæ and procedures. He can be assured, however, that the essential principles involved are really very simple, at least as far as their practical bearings

are concerned. A careful reading of the text need not leave any intelligent person, however innocent of statistical procedure he may be, greatly in the dark with reference to the real significance of facts presented. From the language of statistics there is no escape if we wish to go beyond the limits of personal opinion and individual bias. Worth-while evaluations in higher education will continue to be as rare as they now unhappily are until the rank and file of college and university teachers become able to think in more exact quantitative terms than they are yet accustomed to. Reliability coefficients of .80 to .95 for the new type content examination, as compared with .40 to .60 for the old style essay examination; a correlation of .98 between two scorings of the same new type examination papers, as compared with .60 or .70 for separate gradings of old style essay examinations; correlations of .70 between college success and Thorndike intelligence scores, as compared with .45 between college success and four years of high school marks; a correlation of .46 between scores on a 30-minute section of the Thorndike test and two years of college English marks, as compared with .36 between first semester English marks and second semester English marks;—these are facts of a kind whose significance does not need to be supported by argument.

It may be well to point out that the value of this book does not hinge on the assumption that intelligence can be measured with perfect accuracy, or that the psychology of intelligence is thoroughly understood. As far as the main purposes of the book are concerned the author could without loss have eliminated the term "intelligence" altogether, as carrying implications likely to confuse the issue. By keeping the term he has merely avoided circumlocutions and unnecessary word coinage.

Those who prefer to do so may think of intelligence tests simply as a more or less satisfactory method of arriving at a prediction of an individual's probable success in certain kinds of intellectual work. The value of the prediction may be judged wholly in terms of correlation coefficients or other quantitative, objective evidence. Personal opinion as to what intelligence is or is not may be left out of the question altogether.

The statistical treatment of mental test data has answered once for all the question regarding their validity. Time and again it has been shown that the scores on a single intelligence examination enable us to predict college success as accurately as we can predict it from four years of high school marks. The great value of the intelligence test lies in the fact that it furnishes data not duplicated from any other source. It gives a new line on the student. More than any other kind of information it tells us what grade of work we have a *right* to expect. It gives a favorable starting point for investigating the causes of failures. It enables us to discriminate between the intelligent student whose failure is avoidable and the student whose inferior native ability renders him comparatively non-educable.

The financial aspects of this problem are by no means unworthy of consideration. One university of about 2,500 students and 1½ million annual income spends more than a quarter of a million annually on students who are on probation or who will later be disqualified for poor scholarship. Actual test shows that about 80 per cent of the failing students in this university have a grade of intelligence which renders their failure in such a university practically inevitable.

However, there is no intention to suggest that only the most brilliant young men and women should attend

college. There are colleges and colleges, also departments and departments within the same college. If we had an intelligence index representing the average mental ability in each college, the indices for a hundred representative colleges would be found to cover such an enormous range that the average as well as the brilliant youth could readily select a college that would correspond to his own grade of ability. The purpose of intelligence tests is not to deprive any one of any educational opportunity from which he is fitted by ability to derive normal profit, but rather to enable us to select the type of curriculum from which a given individual can profit, whether he be bright or dull.

It can hardly be denied, however, that the value of a university's product is determined as much by the original quality of the raw material with which it works as it is by salary budget, instructional method or curriculum. It is true that the relatively dull student may and perhaps usually does obtain some benefit from a college education. But a university may "educate" any number of such students without making any appreciable contribution to the advancement of higher learning. Had space permitted, the author of this book could have quoted extensive data from other investigations besides his own showing how greatly the institutions of higher learning in this country differ from one another with respect to the average intellectual quality of their students. Two universities may be regarded by the general public as in all respects equal when an amount of ability below the minimum requisite for bare survival in one would considerably exceed the average in the other. Comparatively speaking, some student bodies are made up largely of intellectual cream while others have drawn almost entirely from the lower levels. Such differences may have been

suspected in some quarters, but only recently has it been possible to measure them.

Just as colleges differ with respect to the intellectual ability of their students, so do the coördinate departments and classes differ within a given college. Generally speaking, students who major in Latin, philosophy, law, mathematics, physics, or chemistry rate higher in intelligence than students electing dentistry, agriculture, household science, art, or music. Of two professors teaching the same subjects in a given department, one may tend to attract bright students while the other attracts mainly the mediocre or inferior.

At Ohio State University the median army Alpha intelligence score of students in the Liberal Arts department was 147; of students in agriculture, 133; of dental students, 115. At the University of Illinois the median Alpha score of law students was 163; of students in agriculture, 139; of music students, 121. At Miami University the median intelligence score necessary to earn the grade A was 162 in philosophy classes, 155 in chemistry, 145 in history, 132 in drawing or music, and 126 in home economics.

For lack of the kinds of facts which tests give us regarding the raw material in our institutions of higher education it has been impossible to evaluate the institutions themselves, their departments, their professors, their curricula, or their methods. If classically trained students were found to make better records in medical school than medical students who had been trained in zoölogy, the fact was explained in terms of the superior mental discipline afforded by the classics. We now know that in some schools the intellectual differences alone are sufficient to account for such a finding.

There are but two essential factors going to make a

university, or for that matter any other kind of school; namely, raw material and educational processes. Each of these factors enters into every sample of the school's product. As long as the raw material is an unknown quantity the merits of the processes also remain unknown and unknowable. Wherever schools are conducted without the use of modern methods of testing the quality of their material, their work is done largely in the dark.

The possible uses of psychological tests as a method of determining fitness for college work in general or fitness to pursue particular courses are especially interesting. It is not yet clear how far it will be found safe to go in omitting various content requirements in the case of students of very superior native ability. Except in certain types of courses it is probable that scholastic success depends far more upon native ability than upon the extent and kind of previous training. A Stanford student whose pre-university training was limited to the completion of the fifth elementary grade has been able, nevertheless, to carry his university work through two years with a record considerably above average. Other cases of this kind have recently been reported from various universities. Certainly a college is justified in permitting the exceptionally able candidate who is short in some of the usual academic requirements to enter by the test route. Properly safeguarded, the plan involves no risk whatever of lowering academic standards. Instead, it puts the emphasis on ability, where it belongs. The candidate who can earn an exceptionally high test score in spite of inadequate training is the best possible bet as regards scholastic promise. This plan of conceding to intellect has the additional advantage that it tends to attract the best and discourage the weakest from entering.

An important function of the university is to locate students of exceptional intellectual superiority who ought to be encouraged to enter the field of research or university teaching. At present we have no accurate knowledge about the average quality of these recruits to the officer ranks of higher education. No one knows whether the average is rising or falling. Some persons in a position to entertain an opinion fear that the worst is happening. It is certain that a continual search for talent carried on with the aid of psychological methods, coupled with suitable vocational counsel and liberal encouragement in the way of assistantships and fellowships, would eventually improve the situation. The future of higher education and of science hangs in no small degree upon the success with which this issue is met.

Another of the important uses of test results is in the sectioning of classes into groups of more homogeneous ability. It need not be argued that test scores should be made the sole basis for such grouping, although they are unquestionably the most important single factor. Our public schools are rapidly adopting the practice of classifying their pupils according to ability. The advantages are so great that once the plan has been tried it is not likely to be abandoned. We may confidently look forward to the time when universities too will cease their present useless attempts to drive the tortoise and the hare abreast.

Although the author rightly gives first place to his data on intelligence tests, the chapters which deal with the improved type of subject-matter examinations are hardly less significant. It is obvious that the problem of measuring our raw material and that of measuring the influence of our educational processes upon this material are inextricably bound together. Our educational

bookkeeping can never be made honest until our present unreliable and subjective examination methods have been replaced by objective tests of high reliability; and when our educational bookkeeping has been made honest its worth-whileness will be more generally appreciated than it now is.

A careful study of this book leads inevitably to the conclusion that it is time for our institutions of higher education to undertake a serious type of personnel research. The results of such research will in the end help to safeguard both the interests of students and the resources of the institution adopting the plan. The personnel bureau has already proved its worth in many large industrial establishments, and it is likely to prove of still greater value in the study of the far more important material with which colleges and universities have to work. The functions of such a bureau would include the following:

1. Preparing and administering tests of general intelligence, and analyzing the results so that they could be used in the study of failures, in the comparison of student social groups and student ability in the various departments, in the vocational guidance of students, and in the selection of promising students for scholarships and assistantships or for encouragement toward advanced study.

2. To conduct research with tests of special aptitude, as in mathematics, science, literature, mechanical engineering, music, art, etc. In one of these lines (music) reasonably successful tests of native ability have already been devised. This entire field is without doubt one of the most promising in applied psychology.

3. To work out, with the assistance of teachers in the various departments concerned, "achievement" tests for

determining the progress students are making in mastering the subject matter of their courses. Tests of this kind are probably feasible in a majority of university subjects and they should in time largely replace the usual examinations, as they are already rapidly doing in elementary and high schools.

4. To make a systematic study of the methods of securing trait ratings of students by their instructors. The first problem here would be to ascertain by experiment what human traits are most amenable to accurate rating. The next task would be to find out what traits correlate highest with success in different academic and vocational lines.

5. To coöperate with the appointment office by supplying certain kinds of data necessary for making intelligent recommendations for positions. Such data would include trait ratings, intelligence scores, and indications of special aptitudes. Recommendations which were known to be based upon a painstaking analysis of students' abilities would soon come to have far greater weight with prospective employers than recommendations based upon off-hand impressions and expressed in the usual unstandardized terms.

6. Other lines of investigation would include such problems as the relation of interests to ability, the relation of high-school success to success in life, etc.

A reasonable sum for the maintenance of the personnel bureau in a college or university would be two per cent of the annual budget for instructional purposes. In a large institution work of real value could be accomplished with one per cent of the instructional budget, but three per cent could hardly be regarded as unreasonable in view of the importance of the issues involved.

CHAPTER TWO

The Columbia College Experiment

In 1919 the Office of Admissions of Columbia University inaugurated an experiment for the purpose of discovering the value of the Thorndike Intelligence Examination for High School Graduates as a criterion for admission to Columbia College. From an attitude of healthy, if not severe, skepticism toward the use of the intelligence test for this purpose, the whole college administration came, within the space of two years, to consider the intelligence test as an indispensable part, not only of the admission machinery, but also of the administration of the college in the Dean's Office. This book represents in part an effort to set forth as briefly as possible the salient facts which brought about this change in attitude. It is hoped that in this process may be presented the major issues concerning the use of more exact measurements of intelligence and of academic achievement from the viewpoint of the administration of the modern college.

The purpose of the present chapter is to set forth the official status of intelligence tests in Columbia College from the beginning of their use to the present time. The treatment will be general, and mainly historical, detailed explanations of theory and practice being reserved for later chapters; the material will consist largely of direct quotations and condensations of excerpts from the official reports of the Dean and of the Director of Admissions of Columbia College.

While the detailed facts that will be thus described occurred in a particular college, the application of the principles involved is obviously not restricted to that

particular college or to any class of colleges or institutions of higher learning. The reason for presenting these facts is that they represent the experience of a typical modern college, and that therefore all institutions of similar description may in some degree profit by that experience. Indeed, the experiment conducted at Columbia College is hardly less illustrative of what might profitably be undertaken by any junior college or high school in America, England, Scotland, Australia, or New Zealand.

The same generalization may be applied to the particular intelligence test described in this book. For the purposes of this volume, both Columbia College and the Thorndike Intelligence Test may be considered as vehicles for describing certain general principles and methods which are equally applicable in any college or high school in the English speaking world. In order to give to the reader an adequate sense of the reality of the facts described, the author has thought best to identify this report rather minutely with Columbia College and with the Thorndike Intelligence Examination, confidently hoping that through the frequent and specific references to a particular college and to a particular test the reader may be brought to see more clearly and certainly than otherwise that the fundamental principles are equally valid for any institution of higher education and for any good intelligence test.

OFFICIAL STATUS OF INTELLIGENCE TESTS IN COLUMBIA COLLEGE

The first significant reference to intelligence tests which we find in the official records of Columbia College is in the Report of the Director of Admissions for the aca-

demic year ending June 30, 1918. Writing in that year and month under the caption "Future Changes," the Director of Admissions expresses in no uncertain terms the growing dissatisfaction with, and the strong feeling of the inadequacy of, the existing methods of admission. In a prophetic passage which has been more than fulfilled, Professor Jones declares:

> It seems highly probable that the conditions in colleges and technical schools after the war shall have ended may call for still more careful administration of entrance requirements, and for modifications which will look toward the rejection of all who are not clearly fitted for the work which they wish to undertake. Economy of the funds to be devoted to education and economy of the student's time and energy will forbid his admission to a course of study for which he is not fitted. We have, of course, endeavored to preclude the unfit in the past, but the admitted impossibility of having completely accurate measures of the intelligence, capacity, and future possibilities of students has seemed to warrant our admitting a few whose fitness was open to some doubt. A good proportion of these have proved their right to a college training, but some have not.

Here we see clearly recognized the importance of sheer intellectual capacity as distinguished from academic preparation, character, industry, interests, health, etc. But, while recognizing the need for an independent measure of mental capacity, existing instruments for measuring the intelligence of college students were regretfully rejected as inadequate:

> Possibly new tests may be devised which will help in solving the problem. The tests for general intelligence so widely discussed for several years past and more recently used in the army cantonments have not yet proved themselves to be sufficiently precise to be of service in solving our problem; but further experience may show how they may be used.

The quality of the hope expressed in the last sentence may be judged from the tendency to place all reliance for immediate improvement on other types of measurements. Quoting again from the same report:

> In any case, a more thoroughgoing employment of the evidences (based on personal opinion) of intellectual and moral competence as a supplement to the formal tests offers some promise of help. The great difficulty, particularly in the case of students coming from large schools, is to obtain such information. No one in a large city high school knows the individual student and his possibilities as the Dean of Columbia College knows the undergraduate, though in some schools the instructors in charge of those planning to go to college do have remarkably accurate and clear knowledge of each student. This is a matter to which schools should be urged to give very careful attention. What the student does out of class and among his fellows is quite as important as his school record in its bearing upon his fitness for a college education.

Such was the status of intelligence tests in June, 1918. The mildness of the hopes entertained as to their possible usefulness is the most striking note. But in the fall of 1918 the War Department invaded the halls of Columbia College along with those of many other American colleges with the Students' Army Training Corps organizations, and incidentally changed the whole outlook there with regard to intelligence tests. The pressure put upon the administration of the college, particularly upon the Admissions Office, was so great that every device about which even the slightest hopes were entertained was seized upon and given a chance to prove its value experimentally.

The regulation establishing the S.A.T.C. placed most of the administrative burden of these officer-training institutions upon the local college authorities, and gave

them very wide discretionary powers. Although the regulation made any candidate eligible for entrance (a) who had had a standard four-year high school course *or its equivalent*, and (b) who could pass a medical examination, other important considerations in the choice of future officers in the army, such as personality and mental capacity, were left entirely in the discretion of the Admissions Office. As an aid in meeting this problem the Director of Admissions resorted to a cautious use of the Thorndike Test of Mental Alertness, with the result that, according to "the universal testimony of those who gave instruction in the Columbia S.A.T.C.," the students so admitted were "exceptionally alert and intelligent."

The short life of the S.A.T.C., the enormous pressure under which the work had to be done, and the consequent incompleteness of records, made exact statistical studies impossible; but the usefulness of the objective mental test was so apparent to all concerned that the Director of Admissions "began at once to consider the possibility of modifying the method of admission to regular college work." When, about the same time, the success of the Army Alpha mental tests in the cantonments became known, the decision was made to ask for appropriate college legislation. During the first months of 1919, then, "the Faculty of Columbia College, after due consideration, adopted a new alternative method of admission in which the mental or intelligence test was the most novel and outstanding feature." (Report of Director of Admissions, June, 1919.)

The significance of this innovation becomes more apparent when we consider, as typical of most colleges, the history of admissions in Columbia College for several years back. Since 1910 candidates had been required

not only to pass the content entrance examinations, but also to submit satisfactory

(*a*) secondary school records,
(*b*) health records,
(*c*) estimates of character and personality.

The new alternative method of admission voted by the faculty in 1919 does not dispense with these three classes of information. The intelligence test is not a substitute for, but an addition to, the method used before 1919. Far from minimizing the influence of school records and personal estimates, the effect of the addition of the intelligence test has been to amplify the details in such records, and to make the use of them more careful and the interpretation of them more exact. The system voted by the faculty in 1919 was based upon the theory that fitness for college work is determined by four elements, each of which is a *sine qua non:*

(1) Preparation,
(2) Character (industry, habits, interests, etc.),
(3) Health,
(4) Intelligence.

It is important to notice that admission under this new system rests upon no one of these four factors alone.[1]

1 That each candidate who is admitted must possess a certain minimum of each of these four qualities is made clear in the report last quoted. As evidence of preparation the candidate must present a record from an acceptable secondary school, showing grades at least as high as those required by the school for certification, in all the subjects required for admission to college; the evidences of moral character and promise as presented in the Principal's recommendation and in at least three personal references must be satisfactory; the candidate's detailed health record must be acceptable to the college medical inspectors. Only if a candidate's records are acceptable in all these respects, is he permitted upon application to substitute the intelligence test for the content entrance examinations. All other entering students take the intelligence test for purposes of record.

The introduction of the intelligence test into the admission machinery represents a large and significant change as to means, but not as to theory. For several years the importance of *capacity* or *power* as opposed to *achievement* or *preparation* had been forcing itself upon college administrators and examining agencies. Before 1910 the emphasis was almost wholly upon *preparation* as measured by the traditional content examinations; and each college was very much a law unto itself as to the number and character of the separate examinations which a candidate must "pass" in order to be eligible. The College Entrance Examination Board began its excellent work of reducing the number and standardizing the character of such content examinations long before 1910, but only in recent years has the Board made a special point of measuring mental ability. The most distinctive step in this direction taken by the Board was the recent adoption of an alternative examination which has more or less descriptively been called "comprehensive." The peculiar merit ascribed to the Comprehensive Examination was that it measured mental capacity as well as content. About the same time the doctrine that the mastery of secondary school content did not necessarily imply ability to achieve success in college became current among college administrators.[1]

Thus, the adoption of the intelligence test in 1919 came about not in answer to a new policy, but as a refinement in the administration of a policy which had been clearly established for some years.

The operation of the "New Method of Admission"

1 "Many a student does and should graduate from high school without being a suitable subject for a college education, just as many a student graduates and should graduate from college who is not of Ph.D. caliber." —Report of Director of Admissions, Columbia University, June, 1919.

during the academic year 1919-20 led the Director of Admissions to say in his report for that year:

> By far the most significant group of students, so far as our admission system is concerned, was the group entering by the new method. . . . That they were a picked group is evident not only from the records which they presented for admission, but also from the records which they made in college. . . . Most of those who failed were students whose scores in the mental test were relatively low, but whose cases seemed to possess sufficient merit to warrant their being given a trial. . . . There were relatively few with high intelligence scores whose records in college were not wholly satisfactory. Careful examination showed that their failure to make first-rate records was due not to lack of intelligence nor to faulty preparation, but to a failure to divide their time and energies properly among the many demands which come to the college student.

Professor Jones gave in the report just quoted summaries of several statistical studies, all indicating a close relationship between intelligence test scores and college standing. Since the evidence will be presented in detail in succeeding chapters, no statistical results will be cited here. But it is interesting to note that after only one year of actual experience with the intelligence test the Director of Admissions felt justified in defending it against the hasty inferences which might be drawn against it on the basis of its failure to correlate better than $r = 0.65$ with success in college. He wrote in the report last quoted:

> It should be remembered that there are many factors other than intelligence which determine a student's standing, and that the psychological examination (or mental test) is not supposed to measure them. . . . There is no thought of using the intelligence test without the student's complete previous record. . . . Occasionally a student of especial promise with

a record which is doubtful in certain particulars may be allowed to take the intelligence examination with the requirement that he pass with a very high mark.

It is hoped that the official status of the intelligence test in Columbia College has been made clear. Considerable space has been given to this exposition because an understanding of just how the test has been applied and used in actual practice is indispensable for an intelligent comprehension of the remainder of this study. For the sake of giving as complete a picture as possible, this chapter is brought to a close with an excerpt from the reports of the Dean of Columbia College. In this extract, written in June, 1920, Dean Herbert E. Hawkes summarizes with his usual directness and clarity the most important aspects of the New Method of Admission as viewed from the Dean's Office:

The correlation between the work of the entire Freshman year for the students who entered by the New Plan and their marks on the mental test is 0.65. The most reliable data available indicate that the highest correlation that can be expected between the work of the Freshman year and the results of the usual college entrance examinations is about 0.45 to 0.55. . . . Every indication points to the mental test as a most useful addition to our machinery of admission. It must be remembered that the group of students who are admitted to college under the New Plan are very carefully winnowed before they are authorized to take the mental test. The correlations obtained, therefore, should be interpreted as referring to the New Plan as a whole rather than to the mental test alone.

In addition to the use of the mental test for admission purposes, it has been most helpful in my office as an aid in arriving at a diagnosis of academic maladies. . . . In several cases the mental test has afforded the clew which has enabled my office, in coöperation with the University Physician, so to

advise the boy that he has not only escaped being dropped, but has become an excellent academic citizen.

The wise use of a new instrument like the mental test requires constant caution and scrupulous checking, but its apparent possibilities for usefulness are so fundamental and far-reaching that a careful and scientific study of its significance is one of the important tasks of the next few years.

CHAPTER THREE

Description of the Thorndike Examination

It is necessary here to devote a chapter to a general characterization of the Thorndike Examination, including a brief historical orientation, a description of certain special features, and an exposition of typical results. Readers who desire at once a more concrete idea of the nature of the examination than can be gained from this chapter are referred to Appendix I, which gives sample questions from all the tests and abbreviated directions for giving and scoring.

ORIGIN OF THE EXAMINATION

Some of the tests in the Thorndike Examination have a history extending back to 1900 or earlier. The opposites test, in one form or another, has given uniformly high correlations with reliable independent measures of intelligence for many years. The completion test, in which omitted words in sentences are supplied by the examinee, has been used successfully since 1895 in the measurement of intellectual capacity. The high relationship between speed in doing simple tasks and ability to do difficult and complex tasks has been conclusively demonstrated during ten years of careful experimentation by a dozen different investigators. But the use of these tests in such combinations as the Thorndike College Entrance Intelligence Examination is more recent.

The first such combination of tests was a battery put together by Professor Thorndike in 1914 at the request of the Carnegie Foundation. The purpose was to secure "a trial set of tests for students entering schools of engi-

neering which might be standardized in respect to form, difficulty, administration, and scoring, and which might be more a measure of the pupil's future promise and less a measure of his previous educational advantages than the ordinary entrance examinations." Bulletin II of the Carnegie Foundation gives an account of the use of this battery of tests with freshmen at Cincinnati, Columbia, and the Massachusetts Institute of Technology. A general notion of the success of the trial set of tests may be gained from one of the often-quoted paragraphs from this Bulletin:

> The records of the thirty-four men tested at Columbia have been followed for three years. Five of the seven who stood highest in the test have received general honors, while four of the seven lowest in the test failed in more than half of the work and left school. The top seven all made more than 125 credits in three years, the middle seven averaged ninety-two credits each in three years, and of the lowest seven the two who did not leave averaged fifty-six credits in three years.

The next step in the development of the College Entrance Intelligence Examination was taken during the War, and resulted in the production, by a series of adaptations, of what is known now as Part I of the examination. At the request of the chief of the American Air Service, the Committee on Classification of Personnel in the Army, under the Chairmanship of Professor E. L. Thorndike, undertook the task of constructing an examination for candidates for admission to the School of Military Aeronautics. The original ten-test Army "Examination *a*," devised by the Yerkes Committee of the National Research Council, was adapted by the elimination of the memory for digits test, by the addition of new and more difficult elements to the other nine tests, and by the introduction of four new tests. These four new

tests were Arithmetic Operations, Syllogisms, Recognition of Identities, and Marking Absurdities. It was necessary to make the Air Service examination more difficult than the Army Alpha because the latter was designed to test all levels of intelligence from Kitchen Police Details to General Staff Officers, while the former was designed only for the highly selected men who were accepted as candidates for commissions in the Air Service. That the examination was successful is apparent from the statement of the military authorities concerned to the effect that it was an "almost indispensable aid toward the speedy elimination of what might otherwise have been doubtful and tedious cases."

A further adaptation, which has already been mentioned in Chapter Two, was made in the fall of 1918 when the Air Service Intelligence Examination was modified by Professor Thorndike at the request of the Director of Admissions of Columbia University for the special requirements of the difficult situation created by the establishment of the S.A.T.C. The modification consisted mainly in shortening the examination, since it was necessary to examine some ten thousand candidates in a very short time. The main contribution of this episode was not so much in the influence it had on the examination as on the Director of Admissions and on the Faculty of Columbia College. It was the intimate successful experience with the examination in selecting students for the S.A.T.C. that contributed, more than any other single factor, to the adoption early in 1919 of the intelligence examination as an alternative method of admission to regular college classes.

Such is the history, in brief, of Part I of the Thorndike College Entrance Intelligence Examination.[1] Parts

[1] For a sample of the various types of questions and tests in the total

II and III were constructed by Professor Thorndike early in 1919 for the special use of Directors of Admissions in American colleges. The tests of Parts II and III have a much shorter history than most of those in Part I, but not a less favorable history as measures of the general intelligence of high school graduates.

CHARACTERISTICS OF THE THORNDIKE EXAMINATION

In addition to the general features common to all intelligence tests built on sound principles of mental measurement, including standardization of form, of administration, of difficulty, and of scoring, the Thorndike Examination possesses certain special characteristics which may be conveniently described at this point.

1. The most striking difference between the Thorndike Examination and other intelligence tests thus far published is its *non-coachability*. The two features which make the Thorndike Examination non-coachable are a monument to the prolific genius of its author. The first is the number of Forms in which the examination is constructed; and the second, the character of the tests themselves.

(*a*) The word "Form" is here used in a technical sense. There are fifteen Forms of Part I, these being named by the capital letters A to O inclusive; and there are more than fifteen Forms of Parts II and III. These Forms are identical in type and difficulty of tests, but different in specific content. Every Form of Part I, for example, has the opposites test, consisting of twenty pairs of op-

examination see Appendix I, pp. 293 to 302. These samples and directions quoted in the appendix are inserted merely to give the reader a general notion of the character of the examination with which we are concerned. Those who desire full details can secure copies of the complete examination, either as samples or in quantity, by writing to the Bureau of Publications, Teachers College, Columbia University, New York, N. Y.

posites; but each Form has its own twenty pairs of opposites. Each Form of Part II has the reading test, consisting of two paragraphs each with six searching interpretative questions; but each Form has its own two paragraphs, different in content from the paragraphs of all other Forms.

Thus a candidate might spend a month memorizing every question in the September, 1919, examination without being helped appreciably on the January, 1920, examination, because only the form of the questions would be similar. Thorough acquaintance with the forms of the questions might operate through the practice and speed factors to raise an individual's score slightly; but this vitiation would not in any event be serious and might be more than counterbalanced by the distraction of totally new problems made more strange by the inhibitions of over-learned bonds of the old problems. The effect would be minimized by the fact that the tests which count most heavily in the final score do not emphasize the speed element. The time allowance is so liberal that it is nearly always *difficulty* rather than *lack of time* which sets a limit upon an individual's achievement in these tests. A candidate would do little better if he had unlimited time.

The solution of the coaching difficulty is not to try to prevent prospective candidates from familiarizing themselves with the types of questions and tests in the examination, but rather to take every precaution to insure that *all* candidates shall have a free and equal opportunity to acquaint themselves with the examination. One Form ought to be published and advertised widely as the *type* of the examination which will be used; and every prospective candidate ought to be urged to study the forms of the questions. The effect of this device

would be to raise the average score on the examination, but the relative positions of the individuals would be nearly the same, possibly a little more reliably placed, due to the elimination of one unreliability factor—the interpretation of directions.

This method of eliminating unfair coaching is absolutely effective, and requires only that standards or "passing grades" be constantly checked up and revised experimentally. The revision of standards is not a burden added by this anti-coaching device; it is a necessity already imposed by the dynamic character of education and life in general. The study of the practice curves of all forms of tests, now happily begun by various experimenters, is an indispensable next step in the development and application of intelligence tests.

This introduces a feature which is of such importance that it ought to receive separate treatment, but which is described at this point because, in addition to its other functions, it serves to a certain extent the purpose here advocated,—that of acquainting the examinee with the nature of the tests before he actually takes the examination.

This feature is the *Fore-Exercise* or *Practice Form* which is given to the candidates for ten minutes immediately preceding the giving of Part I.[1] No Practice Form is provided for Parts II and III, because in the latter speed counts relatively little, whereas in Part I a slight delay in reading the directions or a slight error in interpreting them may cause the loss of several points of credit on a test with which the candidate is really perfectly able to cope.

The main purposes of the fore-exercise are:

(1) to allay nervousness;

1 See Appendix I, 2, p. 302, Directions for Giving, etc.

(2) to eliminate unreliability due to hasty reading and to misunderstanding of directions;

(3) to eliminate unreliability due to variations in the time taken to read the directions for speed tests;

(4) to insure that scores shall be due mainly to differences in ability to cope with the various tests in the examination, and not due appreciably to differences in ability to master quickly and efficiently just one kind of difficulty—that of reading very exact directions.[1]

In addition, and partly because it achieves these ends, the fore-exercise is a strong factor in making unfair coaching harmless.

(*b*) The second reason for the non-coachability of the Thorndike Examination is the character of the tests which are most heavily weighted in the final score. Coaching on the reading test becomes the most legitimate form of education that could be imagined. If the Thorndike test had no influence but that of stimulating prospective students to learn how to read efficiently, it would be worth its cost many times over. Coaching on the completion test will merely tend to make a candidate do for himself what his secondary school teachers often sweat blood in trying to do for him, that is, to acquire an intimate analytic insight into the "mechanics of meaningfulness" in written English. If coaching on the information test [2] will lead a boy to form the habit of assimilating significant bits of information from diver-

1 It might be argued that the directions to a test are a legitimate part of the test, and that any failure to comprehend them exactly or quickly ought to be penalized heavily. But there are two reasons why it would seem inadvisable to lend support to this argument. First, the units would be too large; and second, the ability to comprehend written directions would thereby receive a disproportionately large weight. If we wish to test the ability to follow written directions, let us make up a test of scaled directions, on the order of Test I in Part I; but let us not make every test in the examination one primarily of that one ability.

2 See Appendix I, p. 294.

sified fields of human experience,—and such a habit is the only form of coaching which could help the candidate on the test in question,—it would be folly not to encourage coaching.

2. A second special characteristic of the Thorndike Examination is that it is designed to measure differences in a small part of the total range of human intellect. In this respect it contrasts most violently with the Army Alpha test which was designed to classify human intelligence into only five divisions: Feeble-minded, Morons, Average, Superior, and Very Superior. The Thorndike Examination classifies the last two groups of the Army Alpha into at least twenty significantly different classes. The test is made so difficult at the upper end that the brightest boys have ample room to display their superior wits. Undistributed scores are impossible at the upper end of the Thorndike Examination, which is a matter of very great importance.

3. Quality, rather than speed, of performance is emphasized. It has already been said that unlimited time allowances would not materially raise the scores obtained on most of the tests with the present time limits. In both halves of Part I, speed of adaptation and of performance count probably more than quality of mental reaction; but in Parts II and III, the speed factor is practically eliminated, and it is these parts that are most heavily weighted in the final score.

4. The Thorndike Examination has a greater variety of tests than any other intelligence examination. There are 13 different kinds of tests in Part I, and 8 in Parts II and III, making 21 different kinds of tasks which the candidate is called upon to perform.

5. It is longer in time and content than any intelligence examination thus far published. The actual work-

ing time of the candidate for the whole examination is two hours and fifty minutes, the total time of the examination being not less than three and one-half hours for a group of between 100 and 500 candidates. The content is correspondingly wide and varied, there being no less than 460 different items of graded complexity to which the examinee is given an opportunity to react. This insures so adequate a sampling of both materials and performances as to reduce chance influences to the vanishing point.

6. Since the examination is designed specifically for graduates of the typical American high school, much of the material of the tests is taken from the high school curriculum. The algebra test in Part II is the clearest example of this feature. This is, directly, a content examination; but far from detracting from the value of the examination for our purposes, it adds to that value. As we shall see later, there is no conflict between intelligence and content examinations; the superior value of the intelligence test is not that it measures something different from content, but that it measures content more completely and accurately than the traditional instruments.[1]

7. Due to the length of the examination in time and content, and to the emphasis upon quality of thinking, endurance is tested as in no other intelligence examina-

1 The writer believes that eventually the two kinds of tests will in practice be conjoined for most practical purposes. All that is needed is to familiarize teachers with the technique of the intelligence test. Our quarrel with the traditional content examinations is not that they measure content only, but that they measure it so provincially and inaccurately.

The algebra test tells us the relative gains of students from nearly equal exposure to a given definite field of learning. It is not a test of special interests because it does not go beyond the fundamental operations of elementary algebra. The information test (Part II, Test 8) constitutes a fairly thorough sampling of material from secondary school life. Geometry, biography, history, music, literature, physics, are some of the specific fields represented.

tion. Perseverance and intellectual initiative are put to severe trial in many of the more difficult problems in which the goal is quite recondite at first thought, resembling a puzzle or *impassé* to all but very superior minds.

8. All the directions for the various tests in the examination are in writing. They are not spoken by the examiner, and are therefore uniform and unvarying so far as their presentation is concerned. Any differences in the comprehension of the directions will be due to differences in reading ability, sagacity, and habits of caution of the candidates.[1]

9. Most of the tests in the examination are objectively scored, and can be scored almost mechanically by any average clerical helper under the supervision of a trained examiner. The reading and completion tests and one or two minor tests require some judgment on the part of the scorer, but by careful use of the keys furnished with the examination blanks, the scoring of these tests becomes almost as objective as that of the opposites or arithmetic reasoning test. A trial was made of this point in the scoring of the series of January, 1920, when the reading and completion tests were scored independently by two graduate students. The correlation between the two sets of scores was almost unity, $r = 0.986$. ($M_1 = M_2$. $S_1 = S_2$, and $N = 92$).

10. One feature of the Thorndike Examination which is open to criticism is the elastic division of the time between the various tests. How much time the examinee shall put on each test is left too largely to his caprice and to chance. This renders analysis of the examination

1 This uniformity in the presentation of the directions is, in general, an undoubted advantage and an administrative necessity. But for the less bright candidates, and for the brilliant but hasty and incautious ones, the printed presentation may become the paramount element in the test. This difficulty can be obviated by the wide publication of the *form* of the examination, as suggested in the discussion of coaching, p. 26.

insecure, because we cannot assume that all the men put the same amount of time on each test. Several students in Columbia College have confided to the writer that they ignored certain tests altogether because they knew or thought they knew that they "had no ability of that kind." The inclusion of such zero scores in correlation tables naturally produces some distortion of facts.

11. In the brief limits at our disposal it is impossible to give a complete inventory of the special characteristics of this examination. Only one more will be mentioned here, the system of penalties in the scoring. The theory underlying the assignment of negative scores to absurd or irrelevant answers is too complex for brief statement, but its practical results can be described.

The warning or threat of negative scores in the directions is not always understood clearly by the examinee, and when comprehended may frequently be forgotten. Even when it is understood and retained it is often rendered inoperative because of the widespread American habit of *taking a chance.*

Negative scores are legitimate and necessary in tests which can be answered correctly in a certain proportion of cases by guessing and in which the systematic and measurable chance influences must be neutralized. But in the reading and completion tests pure chance would very rarely if ever produce the correct response. The giving of negative scores in such tests does not reduce, or compensate for, the influence of chance; it increases that influence and penalizes initiative.

TYPES OF DISTRIBUTIONS

Distributions of Thorndike Examination Total Scores. In the three following tables are presented distributions

which may be taken as typical for the classes of high school graduates attracted by colleges like Columbia and Stanford. The distinction between "candidates for ad-

TABLE 1

Score	O	N	(O-N)
33	1	2	3
37		1	1
41	1		1
45		1	1
49		10	10
53		7	7
57	3	11	14
61	2	18	20
65	6	21	27
69	9	28	37
73	13	29	42
77	20	29	49
81	23	28	51
85	22	32	54
89	17	30	47
93	27	19	46
97	16	15	31
101	14	16	30
105	10	9	19
109	6	4	10
113	2	1	3
117	1	1	2
N	193	312	505
M	86	79	81
sigma	13.6	14.7	14.6

TABLE 1. Distributions of Thorndike Intelligence Examination Total Scores for 505 candidates for admission to Columbia College in September, 1919. The Examination was made up of Forms D and E of Part I, and Form B of Parts II and III. The frequency column headed O relates to Old Plan candidates, all of whom were admitted before they took this examination; they took the intelligence test merely for purposes of record. The column headed N relates to New Plan candidates, not all of whom were admitted. The last column is a combination of the other two. The score-index is the midpoint of each respective step.

mission" and "students" (who have been admitted), must be kept in mind if these distributions are to be interpreted intelligently. Tables 1 and 2 refer to *candidates*, many of whom were not admitted; Table 3 refers to *students* actually in college.

In Table 1 we see that, although men admitted under the Old Plan are as a group superior to the candidates for admission under the New Plan, the two distributions are, on the whole, very similar. The difference in average score in favor of the Old Plan entrants is only about 0.5 (sigma); the variability is about equal in the two groups. At least ninety per cent of these New Plan candidates would have been rejected under the Old Plan; about eighty per cent of them were actually admitted under the New Plan. In the next chapter we shall learn that these entrants, who would have been excluded from college but for the intelligence test, succeeded in college nearly as well as the Old Plan entrants, not only academically but in other college activities as well.

In Table 2 are found distributions of scores for each of the four examinations given at Columbia College in 1920. The Old and New Plan candidates are not separated. The differences in the average scores of the four groups must not be interpreted as due appreciably to variations in the difficulty of the various Forms of the examination. The drop of five or six points in the first September examination is almost certainly due, in large measure, to a real difference in the quality of candidates. By September, 1920, the New Plan had become better known, and a larger proportion of men from distant and ill-equipped school districts were attracted by its liberal interpretation of formal requirements. Many of these candidates interpreted the New Plan too liberally. This fact, plainly apparent in the school and character records

TABLE 2

Examinations given in 1920

Score	Jan. BFCC	June GHEE	Sept. IMDD	Sept. FNCC
20			1	
25				
30			1	
35				
40			3	
45		2	7	2
50	3	3	5	5
55	1		19	2
60	4	6	18	5
65	5	5	31	9
70	14	10	51	10
75	16	8	47	15
80	8	13	35	11
85	14	17	48	5
90	14	13	21	1
95	10	15	13	3
100	6	7	11	
105	3	4	6	
110		1	2	1
115		1	1	
N	98	105	320	69
M	80.7	82.4	75.7	71.7
sigma	12.9	14.6	14.4	12.6

TABLE 2. Distributions of Thorndike Examination Total Scores for four groups of candidates for admission to Columbia College in 1920. The Old and New Plan candidates are not separated here.

(1) 98 candidates who took the examination in January, 1920; the examination was made up of Forms B and F of Part I, and C of Parts II and III.

(2) 105 candidates who took the June, 1920 (GHEE) battery.

(3) 320 candidates who took the first (IMDD) of the two examinations given one week apart in September, 1920.

(4) 69 candidates who took the second (FNCC) examination given in September, 1920. Some of these had already taken the IMDD battery given one week earlier, and took this examination in the hope of improving their low scores on the first September examination. This selective factor accounts in part for the difference in average score between the two September, 1920, groups.

Not all of the candidates in any of these groups were admitted.

The score-index at the left is the midpoint of each respective step.

submitted, is alone sufficient to account for the drop from eighty to seventy-five points as an average score. The average score of seventy-one points on the second September, 1920, examination was expected, since many of the men who had received low scores on the first September examination were allowed to repeat the test in the hope of improving or verifying their score on the first trial. Other minor factors also contributed to make this a selected group, both as to average ability and as to variability. The January and June groups are always composed of a rather homogeneous, ambitious, industrious, and somewhat intellectually selected lot of high school graduates.

The writer is indebted to Professor Terman for the data of Table 3. The examination given at Stanford was slightly, and the writer believes happily, unorthodox in that one-half of Part I was omitted, the score of the other half being multiplied by 2.05 in order to compensate for the omission and retain the standards set up for the whole examination by its author. The lower average score obtained by Stanford students as compared with Columbia College candidates and students is unquestionably due to the selective factor. At Columbia only a certain percentage of candidates can be accommodated, even if they were all ideal subjects as to intellect, preparation and character. At Stanford resources are available to care for a much larger proportion of candidates, and consequently the selection is much less harsh. This fact is also indicated by the greater variability of Stanford men.

Another factor which probably had an appreciable influence in lowering the Stanford scores was the lack of any compelling incentive to put forth the greatest possible effort on the part of the examinees. The men tested

were *students* in good standing in Stanford University, *not candidates* for admission; the examination had nothing of the competitiveness of the Columbia tests.

TABLE 3

Score	First-Year		Second-, Third-, and Fourth-Year	
	Men	Women	Men	Women
20-24.9	1		2	
25-29.9	4		2	
30	3		2	
35	2		2	
40	7	2	2	
45	18	4	7	
50	18	3	4	3
55	18	3	10	3
60	35	3	11	7
65	35	6	18	3
70	36	11	20	4
75	30	11	23	5
80	39	10	17	5
85	23	3	12	2
90	9	3	9	
95	7	1	2	
100	5	1	4	1
105	4		1	
110	1		1	1
115	1		1	
120-124.9			1	
N	294	61	151	34
M	70.5	72.3	72.4	72.1
sigma	16.7	13.6	17.7	13.6

TABLE 3. Distributions of Thorndike Examination Total Scores for men and women students of Stanford University who entered and took the examination in September, 1921. These were all undergraduate students who were registering at Stanford for the first time. The examination given at Stanford differed from that given at Columbia in that only one-half of Part I was given. The score on the half of Part I that was given was multiplied by 2.05. Somewhat more than the usual amount of time was allowed for the student to peruse the fore-exercise.

In some respects the difference in variability between Stanford and Columbia men is more significant than the difference in average ability. Great dispersion creates a serious problem for both administrator and instructor. The classification of men for instructional purposes according to intelligence is no longer a convenience, but an imperative necessity, if serious wastes are to be avoided. The lesser variability of women students is quite in keeping with the usual findings in this regard, but in this case is accentuated by the more rigid selection of women at Stanford owing to the limitation of their number to 500.

Distributions of Thorndike Total Scores of Students Who Stay in Columbia College Two Years. In Tables 1 and 2 we saw distributions of scores of various groups of candidates for admission to Columbia College. Not all of those candidates were admitted. Of those admitted, a few failed and left college within two months; others left at mid-year examinations; still others gave up the academic ghost at the end of the year. At the same time, students were withdrawing at the normal rate for the usual miscellaneous reasons. Table 4 shows distributions of intelligence test scores of all men (except foreigners) admitted to Columbia in September, 1919, who were still in uninterrupted attendance in Columbia at the end of the second session, June, 1921. These are men who have survived two full years; but it must be remembered that some of them have barely survived while others have survived *magna cum laude;* and also that some who have not survived left Columbia in good academic standing.

These distributions may be taken as typical of the kind of men who are likely to succeed in a college with standards of intellect and achievement such as are maintained at Columbia. A comparison of them with Tables 1 and 2 will show that elimination has been most heavy in the

lower scores, with the result that the Mean has been pushed up about 0.7 (sigma), and the variability slightly reduced.

TABLE 4

Score	O	N	(O-N)
50	1		1
53			
56	1		1
59	2	1	3
62	1	1	2
65	2	4	6
68		4	4
71	5	6	11
74	5	9	14
77	7	9	16
80	9	6	15
83	10	9	19
86	12	12	24
89	6	12	18
92	11	7	18
95	13	3	16
98	10	7	17
101	8	9	17
104	6	5	11
107	2	5	11
110	5	1	6
113	6		2
116		1	1
N	122	111	233
M	89	86.2	87.7
sigma	12.9	12.4	12.8

TABLE 4. Distributions of Thorndike Total Scores of men admitted to Columbia College in September, 1919, who remained in college two full years without interruption. The first column, headed O, relates to men admitted under the Old Plan, but who took the examination for purposes of record. The second column, headed N, relates to New Plan entrants who were therefore admitted partly on the basis of their intelligence scores. The third column is a combination of the first two. The score-index is the midpoint of each respective step.

Intelligence Standards in Other Colleges. The possibility of making objective comparisons between the mental characteristics of different student bodies provokes the question, Upon what kinds of brains are the energies of representative colleges being spent? No comparison between the services of two institutions can be valid which does not somehow take into account the answer to this question. For example, it might well be that the average intelligence of those who "pass" in one college is not higher than the average intelligence of those who "fail" in another college. In such a case, the college with low passing standard may have ten times as many registrants as the other college, and still not perform one-tenth the service to higher learning and research that the second college performs. In the first college, it may be that only five per cent of those who "pass" are equal in real intellectual ability to the average student who "passes" in the second college.

Table 5 shows that in institutions holding an equal rank in popular esteem, very large differences exist in the averages and variabilities of the intelligence scores secured by their several student bodies.

These differences are so large in some cases as to show beyond question that the caution with which some colleges accept the credentials of students transferring from other colleges is by no means superfluous.

The data show that the average intelligence test score officially reported for first-year men who "pass" their first-semester work in the undergraduate college of one of the two largest universities in America (in point of number of registrants per annum) is barely equal to the average score officially reported for first-year men who "fail" their first semester work in the undergraduate college of the other of those two largest universities.

TABLE 5

Means and Sigmas (approx.) of Thorndike Intelligence Scores of candidates for admission to, and of students in, nine institutions of higher learning in the United States and one preparatory school in Canada. (From official reports of the institutions named.)

	M	sigma	N
1. Columbia College (men)			
Candidates for admission, 1919.....	79	14.5	505
Freshmen admitted, 1919.....	81	13.3	350
Students surviving two years.....	88	12.8	233
Academic Failures (1919-1922).....	69	8.5	90
2. Brown University, Freshmen, 1919....	76	15.0	300
3. Women's College, Eastern State, Freshmen	74	13.3	200(?)
4. Pomona College, California			
Freshmen, men	74	14.5	102
women	71	12.2	106
5. Stanford University, entrants, 1921...			
Freshmen, men	71	16.7	294
women	72	13.6	61
Other Under-Graduates, men	72	17.7	151
women	72	13.6	34
6. Collegiate Institute, Winnipeg, Canada, 1921 graduates	68	13.0	88
7. University of California, Freshmen			
All men	68	15.0	157
Men passing first semester	70	15.0	131
Men failing first semester	61	15.0	26
Women passing first semester	65	11.5	104
Women failing first semester	48	11.0	12
8. Women's College, Southern State, Freshmen	66	13.2	100(?)
9. Trenton, N. J., State Normal School, (Freshmen) Women	64	13.0	170
10. Eastern Agricultural and small colleges	52	16.0	400(?)

The data also show that the chief determiner of the kind of intelligence found in a given institution is not to be explained by differences in size, geographical location, sex, or rank of institution. Freshmen in a women's college in an eastern state get an average Thorndike score

of seventy-four, while the average score obtained by women Freshmen in the University of California *who pass their first semester work* is only sixty-five; women Freshmen in Pomona College secure an average score of seventy-one, and women Freshmen in a New Jersey State Normal School get an average score of sixty-four.

The average score of men who fail during the first semester in the University of California is sixty-one, while that of the women who fail is forty-eight,—a difference of thirteen points, or about 0.85 (sigma). Possibly a third of this difference is accounted for by the fact that the Thorndike Examination is somewhat unfair to women.

The variabilities exhibited by these groups of students are quite interesting. The smallest sigma shown is that of the academic failures in Columbia College, 8.5 points. This indicates that most of the failures are genuinely due to poor intellect, and that good intellects that come into Columbia College are not often allowed to go to waste.

By using these sigmas in connection with the Means, it would appear that less than fifteen per cent of University of California Freshmen reach or exceed the Mean intelligence score of men who survive two years in Columbia College.

STABILITY OF THE MEAN AND S.D. OF THE THORNDIKE EXAMINATION

The stability of the Mean of an examination may be measured by giving two Forms to the same group of individuals and computing the average score on each Form. In order to equalize the effect of practice, the two Forms are given *in rotation*, that is, one-half the group takes Form A first, and the other half of the group takes Form B first. In this fashion each Form receives

equal practice effects. If there are three Forms to compare as to difficulty, one-third of the examinees takes Form A first, one-third takes Form B first, and one-third takes Form C first; if there are four Forms to compare, each Form comes first for only one-fourth of the group.

Unfortunately, this has not been done with college students in sufficient numbers to make the results worthy of detailed presentation. Professor Thorndike studied the records of something over a dozen Freshmen who took several Forms of the examination in rotation, and found that the Means varied not more than three points, while the S.D.'s were practically immobile.

The writer is enabled to present a more reliable study of this very important question through the kindness of Professor Thorndike in allowing the use here of his unpublished study of the records of 171 students in an Eastern State Normal School who took two Forms [1] of the complete intelligence examination.

Since these examinations were given one year apart, it is difficult to estimate the influence of any one of the various factors which might conceivably increase or decrease the average score of 171 students. But it would seem safe to assume that the changes in the individuals during the year were small, that their influence on the scores would be small, and that this influence would tend to increase rather than decrease the average score. We should, therefore, on the assumption that the two Forms

1 These students, of whom ninety per cent were women, took the first Form in 1919, and the second Form one year later. The 1919 examination was made up of Forms C and D of Part I, and Form B of Parts II and III; the 1920 examination was made up of Forms E and F of Part I, and C of Parts II and III.

A convenient way of identifying any particular Form of the complete Thorndike Examination is to call it by the four capital letters which distinguish the Forms that make up the three Parts, and append the date on which it was given. The two examinations just described would be called "CDBB Sept. 1919," and "EFCC Sept. 1920." It is never necessary to name the Practice Form.

were equal in difficulty, expect to find the average score on the second examination slightly higher than that of the first. Similarly, we should expect only a small difference in variability, most probably a slight decrease, due to the elimination of spuriously low scores. How nearly our expectations are realized by actual experiment may be seen in Table 6. The average increases only 1.6 points, and the S.D. decreases less than 1.4 points. This is well within the claim of the author of the test that the average variation in difficulty among the various Forms of the examination would not exceed three points.

TABLE 6

Score	Test, 1919 (CDBB)	Retest, 1920 (EFCC)
25	1	
30		
35	2	2
40	3	4
45	12	3
50	17	10
55	20	26
60	27	26
65	22	30
70	25	25
75	20	18
80	7	16
85	9	4
90	2	5
95	3	1
100	1	1
N	171	171
M	63.7	65.3
sigma	13.2	11.8

TABLE 6. Distributions of scores obtained by 171 Freshmen (of whom ninety per cent were women) in an Eastern State Normal School on two Forms of the Thorndike Examination. The first Form (CDBB) was taken in 1919; the second (EFCC), one year later. The score-index is the midpoint of each respective step.

RELIABILITY OF THE THORNDIKE EXAMINATION

Two Forms of an examination may be identical in difficulty and variability, and still be measures of quite different functions. If the two Forms always give an examinee the same relative position in the two distributions we may conclude that they measure practically the same thing and that they measure it very accurately. In measurements so complex as those which deal with mental and social traits it does not always happen that one examinee gets exactly the same relative position on two Forms of an examination. The extent to which agreement is achieved varies considerably from one examination to another, and is usually most advantageously measured by what is called the Reliability Coefficient. This is nothing more than the ordinary Pearson Product-Moment Coefficient of Correlation, $r = Sxy/\sqrt{Sx^2 . Sy^2}$, applied to a special purpose, and given a slightly different name for convenience of identification. The Reliability Coefficient is also called, more descriptively, the "Self-Correlation" of a test or examination.

The correlation between the scores obtained by the same group of students on two Forms of the Thorndike Examination is 0.85. This is not as high as we should like; but those of us who recall reliability coefficients as high as 0.95 or above for such intelligence tests as the Army Alpha and the Terman Group Test, are likely to underestimate the significance of the coefficient here obtained unless we also recall the influence of "range of talent tested" on all coefficients of correlation. Kelley states (J. E. Research, Vol. 3, No. 5, May, 1921), that "To secure a reliability coefficient of 0.40 from a group composed of children in a single grade is probably indicative of greater, not less, reliability than to secure a

TABLE 7

Test Score	Retest Score EFCC, 1920														
CDBB 1919	35	40	45	50	55	60	65	70	75	80	85	90	95	100	f
25	1														1
30															
35	1	1													2
40		1		2											3
45		1	3	2	5		1								12
50		1		3	11	2									17
55				2	5	9	3	1							20
60				1	3	7	9	4	3						27
65					1	5	8	7	1						22
70					1	1	6	6	6	5					25
75						1	2	5	4	6	2				20
80							1	1	2	2		1			7
85						1		1	2	2	1	2			9
90										1		1			2
95											1	1	1		3
100														1	1
	2	4	3	10	26	26	30	25	18	16	4	5	1	1	171

N = 171 sigma x = 11.8 sigma y = 13.15 M x = 65.32 M y = 63.72 r(xy) = 0.85

TABLE 7. Correlation surface of two Forms of the Thorndike Examination given to 171 students in an Eastern State Normal School. About ninety per cent of these students were women. The score-index is the midpoint of each respective step. (Computations not corrected for guessed averages.)

reliability coefficient of 0.90 from a group composed of children from the second to the twelfth grades inclusive. If it is reasonable to assume that in terms of true ability the spread of talent is four times as great in the eleven grades as in a single grade, the correlation in the second case would need to be 0.914 in order to indicate as close a relationship as that shown by a reliability coefficient of 0.40 in the single grade." [1]

It is, unfortunately, not possible to state exactly how many times the S.D. of the scores of a random sampling of human beings would contain the S.D. of the scores of Columbia College students on the Thorndike Examination. But it is safe to say that if the reliability coefficient were calculated for a group of 1,000 boys distributed equally between the four high school grades and Columbia College Freshmen, it would be above 0.95.

However, this undoubtedly conservative estimate does not alter the fact that in studying the problems of collegiate instruction and administration we are not concerned with the whole gamut of human intelligence. There is a definite limit of stupidity and incapacity below which the direct interests of the college do not reach. Any meaningful statement of the reliability of an examination must be in terms of the range of talent with which the college is concerned. All the data available agree in indicating that the Probable Error of Estimate of true Thorndike Scores from scores on one Form is not greater than 3.1 score-points. This means that the chances are equal that any obtained score, say of eighty points, is within about three points of the true score; that is, the chances are even that the true score lies within the limits

1 "The following formula gives the relationship: $\frac{St}{ST}=\frac{\sqrt{r(1-R)}}{\sqrt{R(1-r)}}$ in which S_t and S_T are the S.D.'s of the two groups in terms of true ability, and r and R are the reliability coefficients of the two groups." —*Loc. cit.*

of seventy-seven and eighty-three points. This statement of the probable error of a Thorndike score means something very definite to the Dean of Columbia College, for example, because he knows that the S.D. of his students is about four times as great as that probable error, and that the correlation of such scores with the best criteria of college success is about 67/70, or .95, as great as the reliability coefficient of those best criteria.[1]

The importance of this matter of reliability, aside from the direct use of reliability coefficients in employing intelligence scores as aids in deciding doubtful cases for ad-

1 Strictly, we can draw no inferences from the normal school data that would apply with absolute security to the situation in Columbia College. The normal school distribution differs from that of Columbia mainly in that the Mean of the latter is about eighteen points, roughly one and one-half times their common S.D., higher. With these facts and reservations in mind, we may say that the measures of reliability of the Thorndike Examination from the normal school data in Table 7 are not far different from those which we would obtain from straight college data.

Readers familiar with statistics will recall that one measure of the reliability of a score is the S.D. of the errors of estimate, which is given by the equation, $S_{1\cdot 2}=S\sqrt{1-r2_{12}}$, in which r_{12} is the correlation between two Forms of the examination in question. For the normal school data, this equation gives $S_{1\cdot 2}=6.3$. Multiplying by .6745 we have the Probable Error of Estimate=4.25 score-points.

But this measure tells us only how nearly we may estimate the score on one Form from that on another. Our main interest is to know how far the score on one Form differs from the True Score, the latter being defined as the score which would be given by taking the average of the scores on an indefinitely large number of Forms. Kelley states (op. cit.) that "The Standard Error in estimating True Scores from x_2 scores is equal to $S\sqrt{1-r_{12}}$." Counting S as 11.8, and r_{12} as 0.85, we find that the Standard Error of Estimate of a True Score from a score on a single Form of the examination is 4.61 score-points. Multiplying by .6745 we find the Probable Error of Estimate of a True Score is 3.1 score-points, as noted above in the text. This means that the chances are even that an obtained score, say 75, is within 3.1 points of the true score.

As Kelley has pointed out (op. cit.), the advantage of this last measure of reliability is that it is independent of the range of talent tested, if the examination is equally good throughout all ranges of talent. Except where comparisons between the reliabilities of different examinations are to be made, this is the perfect measure of reliability. Comparisons cannot be made with this measure because of the provincial character of the unit: a Thorndike unit may be as different from an Army Alpha or Terman Group unit as a Réaumur or Centigrade degree is different from a Fahrenheit degree. When comparisons are to be made the P.E.'s should be expressed in terms of the S.D.'s of the respective distributions.

mission, probation, dismissal, etc., is that the reliability coefficients of measures of two variables set a limit upon the correlation between those variables. Kelley has shown (School and Society 4: 34,71, July 1, 8, 1916) that no measures of a variable, however perfect, can correlate with any other measures of that variable (except in spurious chance cases) to an extent greater than $\sqrt{r_{12}}$, in which r_{12} is the reliability coefficient of the measures last named. It follows logically that this restriction applies to the criterion used as well as to the test used; i.e., to both variables in any correlation, whatever they may be called. Therefore, the maximum possible correlation (except by chance) between two variables is the product of the square roots of their reliability coefficients.[1] The significance of this fact for our purposes will be made clear, it is hoped, in the next two chapters.

INDIRECT MEASURES OF RELIABILITY

Correlations between the Parts of the Thorndike Examination afford an indirect and partial measure of its reliability.

TABLE 8

		N	M	sigma	Minutes	2	3	4	5
1	I D	505	110	17.8	30	.817	.67		
2	I E	505	114	14.3	30		.626		
3	II B	505	125	37.4	60			.544	
4	III B	505	62	17.2	40				
5	1 + 2	505	225	31.3	60				
6	3 + 4	505	185	50.0	100				.746

TABLE 8. Intercorrelations of Parts Ia, Ib, II, and III of the Thorndike Examination, based on records of 505 Old and New Plan candidates for admission to Columbia College who took the examination in September, 1919.

1 That is, $\sqrt{r_{12}} \times \sqrt{rCc}$. In the inexact but highly communicative language of the street, this means that a thing cannot agree with other things very much better than it agrees with itself.

The advantage of these data over those from the normal school is that they involve the kind of high school graduates with which the college must deal.[1]

The reliabilities of certain individual tests and groups of tests are shown in Table 9. The first pair of variables consists of tests 3, 4, and 8 of Part I, and is therefore a mathematical group; the second is a language group, in-

TABLE 9

Tests	Time	M	sigma	r
(1) D348	8	23	6.7	.805
E348	8	23	5.8	
(2) D12611	8	34	5.7	.677
E12611	8	34	4.7	
(3) II B 2	8	11	7.0	.566
III B 2	20	16	7.7	
(4) II 1a-III 1b	18	22	5.2	.480
II 1b-III 1a	18	19	5.8	

TABLE 9. Correlations between different Forms of various groups of tests in the Thorndike Examination, based on records of all candidates (505) who took the examination in September, 1919.

cluding tests 1, 2, 6, and 11 of Part I; the other two are, respectively, the completion and reading tests of Parts II and III. The reading test was divided so that each part would contain one paragraph from each of Parts II and III, and thus equalize the time factor.

The low reliability of the reading test is significant in view of the fact that thirty-six minutes are given to read-

1 Table 8 also shows the Means, sigmas, and time allowances for each Part. It is hoped that no one will object to the omission of the P.E.'s in these Tables. The number of cases for each correlation surface is always given. This will not only allow any one seriously interested in the P.E.'s to calculate them very easily, but it also affords certain direct inferences to the competent thinker which are not easily arrived at otherwise.

ing four paragraphs and answering six questions on each. These reliability coefficients will be recalled in connection with the significance of these tests and groups of tests as expressed by their correlations with various criteria of college success. (See Chs. 4 and 5.)

Additional evidence on reliability is afforded by the normal school data given to the writer by Professor Thorndike and referred to in the preceding pages. Table 10 shows the reliability coefficient of each Part of the Thorndike Examination.

TABLE 10

1919	1920	r	sigma–1	sigma–2	Mean–1	Mean–2	N	Minutes
CDBB	EFCC	.85	13.15	11.8	63.7	65.3	171	160
I C	I E	.822	15.80	13.50	95.3	101.0	171	30
I D	I F	.76	14.15	13.50	105.0	107.0	171	30
II B	II C	.576	30.30	27.80	75.3	90.0	171	60
III B	III C	.600	15.40	15.15	40.0	29.0	171	40

TABLE 10. Correlations between corresponding Parts of two Forms of the Thorndike Examination given to 171 normal school students, of whom ninety per cent were women. The first Form, CDBB, was given in 1919; the second, EFCC, in 1920. Variabilities, Means, and time allowances, for each Form, are also shown. (All correlations reported in this book are Pearson Product-Moment Coefficients.)

Parts II and III make the poorest showing, especially Part II. In the writer's opinion this is in part due to the fact that most of these students were women, and of all the parts of the examination, Part II is least adapted to women. In spite of their low reliability, however, Parts II and III have great significance, as is shown by their correlation with our best criteria of college success.

CONDUCT OF THE EXAMINATION

The conduct of the examination in Columbia College is described in the following memorandum, for which the

writer is indebted to Professor A. L. Jones, Director of University Admissions:

A large room with all the chairs on the same level is provided. The man in charge of the examination has his desk on a high platform from which he can easily be seen and heard by all. He is assisted by one proctor for every fifty candidates (of whom there are usually about four hundred in the first September group). All material is carefully prepared and arranged before the hour at which the examination is to begin. The proctors are fully instructed as to their duties.

As the candidate enters the room, he gives to the chief proctor his receipt for his examination fee, and receives in turn two pencils and a card 3 x 5 in size which bears a serial number, a space for his name, the words *Thorndike Intelligence Examination, Series of September, 1921* (for example), and a space for the score on each Part of the examination and on the total examination. He takes his seat, and at the direction of the examiner in charge writes his name on the card and notes the serial number for entry on each section of the examination to be given him later. The cards are then collected.

The examiner in charge makes a brief introductory statement, giving general instructions and mentioning certain considerations designed to allay nervousness on the part of the candidates. The Practice Form is then distributed by the proctors, and is examined by the candidates (for ten minutes). The other sections follow, with brief rest interval allowed between Parts I and II.

The staff of readers is made up of instructors or graduate students in psychology. Full directions regarding the scoring are in their hands. In September, 1921, the results for some five hundred candidates were scored, entered upon the cards mentioned above, and delivered to the Director of Admissions within thirty hours of the close of the examination.

To this account of the mode of administration of the intelligence examination in Columbia College, it seems desirable to add two elements of policy for the future

conduct of the examination, the wisdom of which is clearly indicated by three years' experience:

(1) The doors of the examining room should be closed and bolted before any of the directions are given out, even for the Practice Form;

(2) No non-English-speaking candidates should be admitted to the intelligence examination.

In the past, late-comers have been admitted even after Part I was half done; and the results have been gravely inconvenient. Not only is the tardy one's record vitiated, but many other examinees are disturbed by the flurry, breathless haste, and prolonged whispering of tardy entrants. Such cases also tend to vitiate statistical studies. In 1919 the records of four tardy entrants into the examination room lowered a correlation nearly .034 between intelligence scores and achievement in college.

Practically the same thing happens when non-English speaking candidates are admitted to the examination. In handling so many papers in a hurried manner it is practically impossible to keep track of special cases; and even if by special administrative arrangements that were possible, it would be largely guesswork to attempt to allow for language difficulty or tardy entrance in each individual case.

The intelligence test is a highly specialized instrument, designed mainly to measure the one factor of mental alertness. It does not, unfortunately, measure the intellect directly, but something else which under certain conditions is almost if not quite synonymous with intelligence. This "something else" that is measured in lieu of intelligence is very complex, and includes such elements as achievement in English language, information in specific fields, ability to think in terms of facts and relations in specific fields of information, expressed in a cer-

tain language. The analysis might be carried to much greater detail. But obviously, the intelligence test is a direct measure of certain types of achievement and habit, and measures intellect only when each individual measured has had an equal opportunity in those specific types of achievement, except insofar as inequalities in opportunity are due to inequalities in native intellect.

The Chinese student who was born and reared two thousand miles in the interior of the Celestial Empire, was graduated from a Methodist High School and College in his home Province, and who has spent probably two or three years in as many American colleges, is not a fit subject for taking the intelligence test with which we are dealing. Differences in opportunity to learn English alone would disqualify him; and when we add the differences in general informational background, cultural habits, etc., the disqualification becomes complete and unquestioned.

THE PREVENTION OF CHEATING

Due to the practical non-coachability of the Thorndike Examination, the problem of preventing "cribbing" is reduced to a minimum. Since different Forms of the tests are used in each examination, the probability that any candidate will have seen the material before entering the examining room is so small as to be negligible. There are two dangers, however, which may arise in the case of astute candidates anxious to get into college at all costs and careless of the nature of those costs:

(1) Copying the work of neighbors in the examining room; and

(2) Sending in a "dummy" to take the examination.

The first is, of course, the historic and common re-

course of the academic villain. The ruse is most dangerous when carefully staged, although it may turn either way. If the candidate copied from is really superior, the intellectual thief may raise his apparent score considerably. This is undoubtedly often the case when the plot is pre-arranged; but, on the other hand, the sympathetic would-be helper may be a very stupid person whose supposed intellectual acumen is in reality nothing more than insinuating self-advertisement.

It is easier for the school yard bully to metamorphose himself into an "intellectual bully" than we imagine. Professor Terman has shown how the glib talker and artful bluffer victimizes the average teacher in this matter. There is some evidence that some very high college grades may be due to the ubiquitous expression of superficial observations in a loud voice and interested manner. It is difficult for an honest instructor not to be seduced by the appearance of deep interest on the part of a shallow mind that grasps at this last straw of academic respectability.

If a frightened candidate has contracted an alliance with such a bully, he may, through his misguided faith, suffer as great a spurious reduction in his obtained score as he would have gained if his faith had been more wisely placed.

The bright boy, contrary to general opinion, is on the whole oftener genuinely modest than the moronish boy. It is the unimaginative moron who constantly over-rates himself and whose desire for approval is insatiable. Not only is the bright boy less likely to make himself conspicuous enough to a low grade intellect to be chosen as a companion in sin, but he is less likely to consent to indulge in such careless conduct, even when it seems perfectly safe.

In any event, liaisons of this sort can be countered effectively. Conspirators can be separated by seating them in an arbitrary, pre-arranged order. The admission cards which are given to the candidates when they show their examination fee receipts to the chief proctor, should be numbered as the seats in the examining room, and should be shuffled before being given out to the men. No two examinees should be separated by less than one or two empty seats, or equivalent space. The aisles should be spacious and plentiful in number, so as to permit the free circulation of the proctors throughout the whole of the examination.

The second form of "beating the game," that of sending in a substitute to take the examination, is more difficult to combat unless great care is taken in the preliminaries. The most available detective means is the comparison of signatures. Each candidate should be required to present for admission to the examination room:

(1) his receipt for the examination fee, bearing his signature as signed in the presence of the Director of Admissions, and countersigned by the latter;

(2) his original application blank, or some part of it, bearing his signature and countersigned by the Director of Admissions.

While the chief proctor is examining these, the candidate should be required to sign his seating card, which should be handed to him from the shuffled pile of cards mentioned above. The three signatures can then be compared. The seating card, with signature and seat number plainly visible, should be clipped to the other credentials (1 and 2 above), and the candidate should be instructed to keep these visible, possibly in the empty chair next to him, during the whole of the examination.

In the case of those repeating the examination, the

credentials for admission to the examining room will be the same as for the first examination, except for the addition of the seating card of the first examination. While the proctor examines these credentials for admission to the re-examination, the candidate will sign a new seating card, easily distinguishable (by color, perhaps) from the seating card of the first examination. If all the signatures tally, he is admitted and seated as before. All these preventive measures are capable of being made a routine process, inexpensive and easy to administer.

COST OF THE INTELLIGENCE EXAMINATION

According to figures furnished by the Director of University Admissions, the cost of giving the Thorndike Examination to about 500 candidates in September, 1921, was about $1,000. This includes the cost of all materials and test blanks, proctor service, scoring, and complete reporting of results. Part of this expense was due to the speed with which it was necessary to complete the task —the complete report was in the hands of the Director of Admissions within thirty hours of the close of the examination period.

No exact comparison can be made between the cost of the old and new, or content and intelligence examinations. The College Entrance Examination Board, operating under conditions not essentially different from those which obtain when content entrance examinations are administered by the college, reports the expenditure per candidate in 1919, 1920 and 1921 to have been, respectively, $6.89, $7.26 and $8.08. The Board has also announced, through the Report of the Secretary for 1921, that the examination fee hereafter will be $9.00 per candidate instead of $6.00 as heretofore.

The cost of examining 20,000 candidates at the new rate of the Board would be $180,000. The cost of the intelligence examination for 20,000 candidates at the rate reported by the Director of Admissions of Columbia would be $40,000. The difference is $140,000.

METHODS OF KEEPING RECORDS

Very little can be said about methods of keeping records that would be equally applicable in all collegiate conditions. For some purposes a list of the names in alphabetic order with the intelligence scores is all that is necessary. Such a list is a valuable supplementary record in all cases. From the generalization made in the second chapter that intelligence test results find their greatest usefulness in combination with all other available information, we may generalize here and say that, other things being equal, that form of record is best which brings together in one view all the information that is available about each individual, and which is self-continuing and easily kept up to date. It is unnecessary to add that all scores should be expressed in terms of defined units reckoned from a defined point of reference. The most exact method is to express all scores in terms of S.D. above or below the average of some standard distribution; but this is not always either necessary or best, although it should always be possible to transmute whatever score is used into such terms.

All things considered, the card system seems most advantageous. Contrary to the apparently general desire to use the smallest possible card, our experience indicates that a large card is not only advantageous but necessary. Six by eight or nine inches is recommended, and heavy durable stock. If the card is to be used

mainly or partly for research purposes, spaces ought to be allowed for analyzed scores in addition to spaces for total scores. The Thorndike Examination, for example, has many different kinds of tests which may be grouped in various ways; and college success, depending as it does on various factors, such as discipline, health, attendance, interests, freedom from worry and distractions, intelligence, and achievement in from three to seven different college courses, should not be expressed in one single index only. Only by analysis may we hope to secure a workable index of college success, or a workable series of indices.

Where the cards are to be used for research purposes or for any statistical investigation, flexibility of space allotments becomes an important consideration. We cannot always foresee the various items or indices that may become significant, hence it is advisable to leave a large number of free spaces, as the card on the following page, designed by Professor E. L. Thorndike for the special purpose of the experiment which forms the basis of this book.

The main characteristic of this card is its great flexibility. The free spaces in the health line have already been used for the totally unforeseen morphologic data mentioned in Chapter Four. The fourth line is designed to give a complete picture of the individual's secondary school record and recommendations. The fifth line provides space for his intelligence scores, 1st, 2d, and 3d trials, and for scores on designated groups of tests in the mental test. The 6th, 7th, and 8th lines provide spaces for college scholarship data for eight semesters, with free space for summer school records, etc. The lower third of the card provides sixty free spaces for whatever might develop in the course of the investigation.

NAME last	first	middle	AGE yr. mo.	LAST SCHOOL	FUTURE COURSE	No.
HOME city or town	FATHER'S BIRTHPLACE	MOTHER'S BIRTHPLACE	FATHER'S OCCUPATION	RELIGIOUS CONNECTION		

Health	Eye	Ear	Respir.	Digest.	Nerv.	Nutrit.	Valitity	A	B	C	D	E	F	G	H	I	J

Regents Median of 15 highest	M.S.B.	Compre	Subjects Conditions	Units	Subjects Advanced	Points Standing	Raw Abil 13-10	Corr	Raw Char 456	Corr	Raw Char 789	Corr	High School Marks Median	Raw Eng	Corr Hist	Raw Lang.	Corr	Raw Math.	Corr	Raw Science	Corr

Date TEST	Form	Score	Date RETEST	Form	Score	Date 2ND RETEST	Form	Score	Language	Math.	Information	Reading	Trade								

	F1	F2	S1	S2	J1	J2	Se1	Se2													
Points earned																					
Average highest 12																					
Scholarship Score																					

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54	55	56	57	58	59	60

The writer has found them indispensable. They have been used for various unforeseen items of data, but mainly for records of college achievement by courses and semesters. All that is necessary in the use of these free unnamed spaces is the keeping of a full and up-to-date key, showing exactly just what the figure in each space stands for. These sixty spaces are sufficient for showing in analyzed form the grades of one student in every subject matter in his baccalaureate course.

For the use of the Dean, some of these spaces have been set aside for such items as Probation, Warning, Health, Hours spent in commuting, Hours spent in outside work, Hours spent in college activities (athletics, dramatics, student-journalism, etc.), Name of his Faculty Adviser, Fraternity connections, and many others too numerous to mention. The cards should be kept easily accessible to the Dean, preferably in one of the rooms of his suite.

The value of such a complete and compact record for statistical purposes is perfectly obvious to any reader who has had statistical experience; and its value for the Dean of a college is apparent from the condition which would prevail in its absence. As an example of the latter condition let us take a hypothetical case.

Freshman X creeps into the Dean's Office in a last effort to stave off probation, or some other badge of partial or complete academic failure. He alleges illness, among other excuses, and begs for another chance. The first thing the Dean wishes to know is whether the boy is really failing or is merely frightened, as at least a few conscientious champions of the green card are wont to be just before examination time.

He finds in a huge volume supplied by the Registrar, the boy's last reported grades, which are very unfavor-

able. Perhaps the boy has not yet found himself, academically speaking. His secondary school record may throw some light on the very grave problem of advising a hopeful young human being in the best way; the Director of Admissions is consulted by telephone. A clerk is set on the job of finding the boy's application records; the files are often up to date, so in due time the Director of Admissions reports fair records in high school, well recommended by the school Principal and the local Judge.

While holding the wire for this last bit of information, hasty questions reveal that this is the boy who commutes two hours per day on a very poor and crowded train, works for a living six hours on each of four nights in the week, and is reported by the University Physician to be undernourished. With the centralized card recommended above, all this information and much more would have been secured by the Dean in one glance.

CHAPTER FOUR

Validity of the Thorndike Examination

prediction of success and failure in college

The average intelligence scores of men who "pass" and of those who "fail" in college give us a rough index of the predictive value of the intelligence test. The average score of those who survive one semester in Columbia College is eighty-six, with S.D. of thirteen; while the average of those who survive two years is eighty-eight, with S.D. of 12.4. This means that about eighty-five per cent of the men who survive two years in Columbia College have an intelligence score above seventy-five. This is in violent contrast with the average and S.D. of the men who fail in the first semester, either wholly or in part. Data collected for three fall semesters show that the average score of academic failures in Columbia College is not above sixty-eight, with an S.D. of less than nine points. More than ninety-three per cent of those that pass are above the average of those that fail, and less than two per cent of the latter reach the average intelligence score of the former.

(1) The average intelligence score of those dismissed for poor scholarship at the end of the first semesters in 1919, 1920, and 1921 was, according to data furnished by the Dean of Columbia College, sixty-eight points, with an S.D. of about 8.5 points.

(2) The average of men placed on probation for academic reasons in 1920-1921 was sixty-eight. No probationer during this period had an intelligence score above eighty-four.

(3) The average score of men admitted in September, 1920, who were deficient in English preparation was seventy-one. Several of these men scored about eighty in intelligence, but

they were found to be only formally deficient. They did uniformly good work in college English, while those below sixty-eight failed without exception.

(4) Of about forty-six men with intelligence scores below sixty-eight, exclusive of foreigners, who were admitted in

TABLE 11

Thorndike Intelligence Score	Probationers 1920-1921	Men Deficient in English Preparation 1920-1921
44	1	2
47		
50		
53		2
56	2	2
59	4	4
62	3	4
65	1	3
68	7	9
71	6	
74	4	5
77	2	6
80	2	4
83	2	2
86		1
89		1
92		1
95		2
98		1
N	34	59
M	68	71
sigma	8.4	11.2

TABLE 11. Distributions of intelligence scores of Columbia College students in 1920-1921 who were on probation and who were deficient in English preparation.

September, 1919, thirty-four survived the first semester; twenty-five survived one year; and twelve have survived two years. The success of these twelve men has been barely enough to keep them on the rolls of the college by the exercise of very hopeful discretion on the part of the Dean. Just how badly these men have fared in competition with their more intelligent

fellow-students is apparent in Tables 12 and 13, as well as in other Tables in this chapter.

TABLE 12

Number of men	Thorndike Intelligence	Scholarship Score	Average Quality Index of 64 Hours	Number of Hours of College Work at Each Grade of Merit		
				C	B	A
(1)	(2)	(3)	(4)	(5)	(6)	(7)
7	68—	246	3.84 (D)	41	31	22
12	86	420	6.56 (C)	70	53	38
9	101	500	7.87 (B)	83	63	46
7	105+	624	9.75 (B+)	104	78	57

TABLE 12. The quantity and quality of academic work done in two full years of residence by students in Columbia College who secured Thorndike scores of 68 (and below), 86, 101, and 105+, based on records of 111 men admitted in September, 1919, under the New Plan and who survived two years in college.

The scholarship score is the sum of the products of the grades multiplied by the number of points represented by each grade. The quality grades A, B, C, D, and F are assigned numerical values 11, 8, 6, 4, and 1, respectively. A grade of A in a three-hour course contributes 33 points to the scholarship score; a grade of A in a six-hour course gives 66 points. A student who in one semester gets A in six hours of college work, B in 8, and C in 3, would have a scholarship score for that semester of 148.

A boy taking a normal program of 16 hours and making an average grade of C will have a scholarship score of 96; it will facilitate our thinking to fix in our minds that the average scholarship score for one semester is roughly 100, that for a full year roughly 200, and that for two full years roughly 400.

The four groups here presented may be compared in three ways. The first is by the scholarship score just described. The second is by means of column 4, which shows for each group the average quality index on the basis of a normal 64-hour (16 hours for each of four semesters) program; this quality index is the quotient of the scholarship score divided by 64.

The third method of comparison is afforded by columns 5, 6, and 7, which show in three different but equivalent ways the relative quantity of work done by the four groups. Each of these columns represents the same achievement, but in column 5 it is in terms of C grade work, in column 6 in terms of B grade work, and in column 7 in terms of A grade work.

The seven New Plan men with intelligence scores below sixty-eight have an average S.S.[1] of 246, which is a score that would be earned if a normal program of sixty-four hours were completed with an average grade of D (conditional failure), or if forty-one hours were completed with an average grade of C, or thirty-one hours of grade B, or twenty-two hours of grade A. The most favorable interpretation is that they do only two-thirds of the normal quantity of work, and that at less than average

TABLE 13

(1)	(2)	(3)	(4)	(5)	(6)	(7)
12	68—	330	5.10	55	42	30
18	89	480	7.50	80	60	42
17	101	498	7.78	83	62	46
9	108+	630	9.84	105	79	57

TABLE 13. The same as Table 12, except that these data are based on the records of the 111 men represented in Table 12 combined with the records of 122 Old Plan men admitted in September, 1919, who also survived two full years.

quality. Contrasted with the men of intelligence scores above 105, these seven men present a spectacle well worthy the thoughtful regard of every college administrator. This contrast, in brief, is an average quality index of D — over against one of A, and forty-one over against 104 hours of grade C college work completed in two full years of residence in Columbia College. When it is necessary to limit registration so strictly because of lack of accommodations, it would seem difficult to justify the continued presence of such low grade intellects in a college of high standards. These men of low capacity un-

1 Abbreviation for Scholarship Score.

doubtedly gain something, they probably gain much; but others, if they were admitted, could gain more and would gain more. The same story is told in Table 13.

In Table 14 we find the same story again, told in a negative way. About one-half the men with intelligence scores in the neighborhood of seventy fail in at least one-third of a normal program of college work.

TABLE 14

Intelligence Score	% making F in 5 or more Hours	% making F in 11 or more Hours	% making F or D in 5 or more Hours	% making F or D in 11 or more Hours
64-75	38	18	79	48
76-87	23	6.2	58	35
88-99	12	2.0	45	14
100+	3.2	0.0	19	3

TABLE 14. Percentages of New Plan men admitted in September, 1919, who survived one year in Columbia College. N = 180. The normal number of hours taken in one year is 32; hence, according to the figures above, about one-half the men making intelligence scores between 64 and 75 get grades of F or D in one-third of a normal program. The average number of hours of "failed work" is about half of a normal program.

In judging the significance of these data, we must remember that the men here considered are far from being a random sampling of men who would get seventy on the Thorndike Examination. The men who received scores below seventy-five were hand-picked as regards other qualifications; and of those admitted several dropped out during the year for poor scholarship whose records are not included in these calculations. We must also notice that the men scoring below seventy-five attempt only three-fourths of a normal program, and therefore that they fail in about two-thirds of the work they actually attempt.

All these data show that the man with an intelligence score in the neighborhood of seventy is a bad risk for Columbia College. But there is no definite "passing score" on the intelligence examination. On the contrary we have what might be called a "critical area" above which admission is almost certain as far as intellect is concerned, and below which rejection is similarly certain, but within which the most careful discretion and weighing of other factors

TABLE 15

Intelligence Scores	% making F or D in 7 or more Hours in Two Years
68 and below	83.3
64-75	71.4
76-87	48.6
88-99	29.0
100+	8.0

TABLE 15. Percentages of men at various intelligence levels who make a grade of F or D in 7 or more hours of work during two years of residence in Columbia College. Based on records of 233 men (exclusive of foreigners) admitted under both Old and New Plans in September, 1919. The men below 68 attempt an average of 50 hours in two years, much of which is repeated. Those above 100 attempt an average of 70 hours.

must be the rule. The upper limit of this "danger zone" is certainly not above eighty-five, and probably not above eighty. The lower limit is not below sixty. Men between the limits of sixty-five and seventy-five should be admitted to Columbia College only if they present evidence of extraordinary character and industry, or if there are no better candidates who might be admitted. Ninety-five of every hundred men below sixty-five admitted to Columbia will fail and be dismissed within two years. The admission of such men is not to the best interests of the college, of society, or of the individual. From the standpoint of the

college and of society it is like investing money in one per cent bonds when gilt-edge ten per cent bonds are begging; from the viewpoint of the dullard concerned, it may be almost tragic.

CORRELATION OF INTELLIGENCE SCORES WITH COLLEGE SUCCESS

Since the facts with which we are concerned are in a continuous series and permit no definite dichotomy, we cannot be content with knowing approximately the location and nature of the "passing grade" or "critical area." We wish to know not only whether men with low scores fail often and succeed rarely, but also whether men with high scores fail rarely and succeed frequently and brilliantly. The critical area determined for Columbia College would apply only to Columbia or to colleges which maintain equivalent standards; furthermore, this danger zone would apply only to the Columbia of the present day, since we do not know that present standards are permanent. In short, we wish to know whether there is a general relationship between intelligence scores and achievement in college.

There are various methods of studying this question, not the least significant of which is that represented in Tables 12, 13, 14, and 15. In whatever way we interpret these tables, we see that high scores on the intelligence test mean high achievement in college, and *vice versa;* and that low scores in one mean low scores in the other. In Table 12 we find fairly uniform increments in both quantity and quality of academic work done as we go up the scale of intelligence scores. In terms of C grade work, the men above 105 in intelligence score are within twenty hours of graduation at the end of the second year, while the men

below sixty-eight (although hand-picked) have eighty-three hours of the same quality of work to do before graduation. The minimum requirement for the B.A. degree in Columbia College is 124 hours. In Table 14 we find a progressive decrease in the percentage of men who get F and D grades as we ascend the intelligence scale.

There is little need of piling up evidence of this kind, but it may interest some to see three often-quoted tables

TABLE 16

Intelligence Scores	N	F	D	C	B	A	Total points attempted	Points passed (C, B, A)	Average Scholarship Score
70-72	10	33	15	46	31		125	77	61.7
86-88	10	9	21	53	51	6	140	110	88.5
104-106	10		9	49	89	19	166	157	125.0

TABLE 16. College achievement of each of three groups of ten men selected at random from three different ranges of intelligence scores. Column N indicates the number of men in each group. Columns F, D, C, B, and A show the number of hours of college work completed by each group during the first semester of the Freshman year at each grade of quality. The Scholarship Score is calculated as in Table 12.

drawn up by Mr. Harold K. Chadwick, Assistant Director of Admissions of Columbia University, on the basis of Freshman records at the end of the first semester in the session of 1919-1920.

Ten men selected at random from among those Freshmen who made intelligence scores between seventy and seventy-two attempted a total of 125 hours, or an average of 12.5 hours each. In spite of this very light academic load, these men failed in nearly forty per cent of the work attempted, and made no grade above B. Ten similarly

selected men scoring between eighty-six and eighty-eight attempted 140 hours of college work in the same time;

TABLE 17

Intelligence Scores	Lower Quartile	Median	Upper Quartile
70-75	F	C−	C+
85-90	C−	C	B−
105+	C+	B	B+

TABLE 17. Medians and Quartiles of quality grades of all men admitted in September, 1919, who secured indicated intelligence scores. These are for grades received at the end of the first semester of the Freshman year.

they failed in twenty-two per cent of their work, passed 110 hours, of which six were of A grade. On the average these ten men passed eleven hours as compared with 7.7

TABLE 18

Intelligence Scores	N	Number Averaging B or better	Per cent averaging B or better
60-70	50	2	4
70-80	50	4	8
80-90	50	7	14
90-100	50	15	30
100-117	50	25	50

TABLE 18. Numbers and percentages of men at various intelligence levels who make an average grade of B or better in the first semester in Columbia College. Fifty men were selected at random in each range of the intelligence scale indicated.

hours passed by the men scoring between seventy and seventy-two, and they did this work at a higher grade of quality. Ten men with scores between 104 and 106 at-

tempted in the same time 166 points; they received no F grades, only five per cent D grades, and passed 157 hours of work, of which 108 were of B or A quality.

More exact measures of relationship are afforded by the various correlation coefficients. All the correlations reported in this volume are Pearson Product-Moment Coefficients, given by the mean product of the deviations expressed in terms of the product of the sigmas. In Table 19 are presented various correlations between in-

TABLE 19

Criteria of Success	156 Old Plan Men	199 New Plan Men
Number of credit hours earned	0.337	0.482
Median College Grade	.512	.537
Scholarship Score (points x grades)	.488	.598
Average Grade plus Points Attempted		.580
Average Grade x Points Attempted		.590
Kelley Combination of grades and points		.588

TABLE 19. Correlations of Thorndike Examination Total Scores with indicated indices of college success in the first semester for men admitted under the Old and New Plans in September, 1919.

telligence test scores and various indices of college success in the first semester of the Freshman year.

Our chief interest in this investigation is in the group of men admitted under the New Plan; but for purposes of comparison, some correlations for Old Plan men are also included. The latter are uniformly lower than those for New Plan men. The cause of this difference is complex. The Old Plan men are slightly higher in the scale of intellect, their variability is somewhat less, their academic advantages and experiences have been better. They are selected morally as well as intellectually. Some come to

college not because they are ambitious about drinking deep in the Pierian Spring, but because their parents desire it and their social set expects it. Their thoughts run somewhat to those aspects of collegiate life that are extra-curricular if not extra-mural. Among such men serious discrepancies between capacity and achievement are not surprising. The decisive factor is interest, not ability. Even when interest is favorable to academic activities, the Old Plan men do not feel that it is especially incumbent upon them to put forth their maximal effort.

By contrast, the New Plan men are early and unequivocally impressed with the fact that they are on trial. They know before they arrive that they are going to compete with men who have been exposed to more and probably better educational influences. They are therefore selected as to interests and courage and ambition, and most of them probably put forth their maximal effort most of the time, certainly during the first semester.

It is not to be inferred, however, that factors other than intelligence had no influence on the academic standing of New Plan men at the end of the first semester. The argument of the two preceding paragraphs is simply that the non-intellectual factors acted on the records of the Old Plan men with less restraint than on those of the New Plan men. The highest correlations found here between intelligence scores and first semester college achievement are not above 0.60. At first sight, this will undoubtedly seem unsatisfactorily low to some readers; but when we consider how many and how strong are the influences besides intelligence that affect college work, it is patent that even if we had a "perfect" measure of intellectual capacity, we could not expect perfect correlations with academic standing. Health, industry, interests; freedom from worry or grief; freedom from distractions,

such as voluntary activities like athletics, dances, student journalism, etc., or necessary penances like commuting on crowded trains or working nights for a living; happy choice of and adaptation to particular branches of study —to expect a correlation of 1.00 between intelligence scores and the result of such complex and variable forces as these would, of course, be absurd.

Another factor, about which we can happily be more definite, which limits the maximal possible correlation, is the reliability of our measures. Our intelligence examination agrees with itself in the measurement of the same person or persons to the extent indicated by a self-correlation of 0.85; and the correlation of college success in the first year with that in the second year is about 0.70. By Kelley's theorem quoted above it follows that no test or other instrument of prognosis could correlate higher with scholarship scores than the square root of .70 which is .84. This means that if the Thorndike Intelligence Examination correlated .84 with scholarship scores, it would be proven to be a perfect measure of future scholarship; that is, a perfect instrument of prognosis of scholarship. The correlation of .84, rather than 1.00, should be considered the goal, therefore, so far as the Intelligence Examination is concerned. In considering the facts of Table 19, we must not forget that the various indices of college success there used are based on the work of only one (the first) semester in college.

THE SCHOLARSHIP SCORE

As used in this volume, "scholarship score" is to be defined as in Table 12; i.e., as the sum of the products of (Hours x Grades) for all Hours of college work attempted. The numerical values assigned to letter-grades

F, D, C, B, and A are, respectively, one, four, six, eight, and eleven. If a student is taking eighteen Hours, and makes A in three Hours, B in six, C in six, D in two, and F in one, his scholarship score for the semester would be 126:

Quality of Work	Quantity in Hours	Products
A = 11	3	33
B = 8	6	48
C = 6	6	36
D = 4	2	8
F = 1	1	1
		126

It is believed that this scholarship score, which hereafter will be designated as S.S., is the best index of college success available. The college degree is based on quantity as well as quality of work done. In Columbia College the baccalaureate degree is granted, if all other requirements are satisfied, on the completion of a definite quantity of work, namely, 124 Hours, of a quality not below the grade of C. As far as the writer knows, all colleges insist on a certain minimum of quantity as well as a certain minimum of quality of work. Yet it has been customary to measure college achievement for administrative and disciplinary purposes almost wholly in terms of quality. The injustice of ignoring quantity is apparent in Table 16. The average quality grade of the low group is in the C's, while that of the high group is in the B's; but this is only half the difference, because the low group attempts only seventy-five per cent as much and passes less than fifty per cent as much work as the high group.

As used by Phi Beta Kappa, the pure quality index

is probably nearly always adequate and just, because the quantity of work involved varies little as between the individuals whose admission to the baccalaureate Hall of Fame is at stake. But for questions of internal undergraduate administration, the pure quality index is often gravely inadequate. The prescribed limits for quantity of work for which an undergraduate may register without special permission, are twelve and nineteen Hours. When it is reported that during a given semester two boys each received a median grade of C, the obvious inference is that the boys are about on a par as to college achievement; whereas the truth may be that one received four D's, four C's, and four B's, while the other received six D's, six C's, and six B's.

The Phi Beta Kappa system would be strengthened if it applied at both ends of the quality scale the well-established principle which it recognizes and applies at the lower end. The numerical values given by Phi Beta Kappa to the letter-grades A, B, C, D, and F are, respectively, one, two, three, four, and seven. The difference between F and D is recognized as much greater than the difference between any other consecutive pair of letter-grades above F. But by parity of reasoning, the difference between A and B is as great as, if not greater than, that between F and D. There is no evidence that it is smaller, and there is much that it is greater.

Instructors everywhere are very cautious, if not reluctant, to give the A grade. They give A instead of B + only when there is no doubt at all in their minds about the individual's extraordinary merit. Whenever we find the A grade, then, we may be sure that it indicates ability and achievement not merely above the B grade, but distinctly and undoubtedly above B grade. On both theoretical and experimental grounds it is doubtful whether the step from

D to F is as great as that from B to A. Many instructors are inclined to use the D grade very sparingly, preferring to give F to all men who do not complete the work satisfactorily, and D only to those whose failure is accompanied by fewer manifestations of incapacity or laziness. Also, the chances that a mediocre man may get an A are much less than that he may get a spuriously low grade. In weighting the F three times as heavily as the A grades, chance is given too much influence. All these considerations seem to indicate the superiority of the evaluations

TABLE 20

Correlations between intelligence scores and average grade in college work done in indicated periods of time, based on records of 111 men admitted under the New Plan in September, 1919, who remained in Columbia College two full years.

Thorndike Scores	r	M 88.16	sigma 12.36
1. Freshman Year	0.578	C+	0.5 [1]
2. Sophomore Year	.588	C+	0.5
3. Freshman and Sophomore Years	.622	C+	0.5

[1] Letter-grade.

of the letter-grades used in calculating our scholarship score, or S.S.

That the quantity index is significant, even in total independence of the quality index, is evident from the data of Tables 20 and 21. Its significance increases as the period of time in college increases. The quality index is consistently more significant and also more stable, since its correlation with intelligence scores changes little from one year to another; but this superiority is due to the fact that it is not wholly or purely a quality index.

Indirectly it does take some account of the quantity factor; and the inference is rather clear that its significance would increase if more adequate account of quantity were taken.

Accepting for the moment the theory that the Thorndike Examination is a good measure of intelligence, and that college marks ought to agree closely with true intelligence scores, we may say that the inference as to the importance of quantity is supported by the facts of

TABLE 21

Correlations between intelligence scores and number of Hours of college work done in indicated periods of time. Same population as in Table 20.

	r	M	sigma
1. First Semester in College	0.454	15	4.40
2. Second	.395	15	4.00
3. Third	.395	15	4.12
4. Fourth	.548	15	4.72
5. First Full Year	.491	30	7.40
6. Second	.552	30	7.78
7. Two Full Years	.598	60	12.35

Table 22. If we also accept the theory that the most constant and reliable measure of college work is the best one, then we may say that Tables 39, 40, 41, 42, and 43 in Chapter Six, practically prove that the inference is true, and that our S.S. (Scholarship Score) is the best measure of college success thus far developed. Of course, further experience may refine it.

In Table 22 we have our best picture of the relationship between Thorndike Intelligence Scores and college standing. The correlation between the three-hour intelligence test and the two-year scholarship score is 0.672.

This coefficient of 0.672 may be accepted as extremely

reliable. All the records involved were checked and rechecked to insure impartial accuracy. All cases of late entrance to the intelligence examination room were thrown out; records of all men not native to the English language were excluded; cases in which long illnesses were known to have occurred were excluded. Cases of discipline were excluded only when there could be no doubt as to the vitiation introduced by conduct as opposed to intellect. All cases of voluntary withdrawals were excluded. Finally, only the records of New Plan men were included, so that

TABLE 22

Correlations between intelligence scores and scholarship scores earned in indicated periods of time, based on records of 111 men admitted under the New Plan in September, 1919, who remained in Columbia College two full years.

	r	M	sigma
Freshman Year	0.632	220	52.8
Sophomore Year	.620	210	65.4
Freshman and Sophomore Years	.672	440	109.0

the relationship found is that for men admitted mainly on the basis of the intelligence examination.

One vitiation deserves separate mention. A large percentage of Columbia College students are partly self-supporting, and some earn their entire support while in college. At least three of the men whose records are included in Table 22 work for a living as much as five hours per day six days in each week. Many of them commute, spending from one to four hours per school day on crowded trains. Some of them work intermittently.

No one will deny the strong influence which these facts exert on the correlation between intelligence and achievement in college; nor will any one deny the difficulty, if not

impossibility, of evaluating them numerically. Their effect is to lower the magnitude of the correlation coefficient. Thus, while we cannot evaluate or estimate the amount of the vitiation introduced by the complex influences in question, we at least know that they operate not to exaggerate our correlation index, but to decrease it spuriously.

TABLE 23

Correlations of Thorndike Examination scores with scholarship scores earned in indicated periods of time by 122 men admitted under the Old Plan in September, 1919, who remained in Columbia College two years.

	r	M	sigma
Freshman Year	0.514	240	55.0
Sophomore Year	.373	240	61.4
Freshman and Sophomore Years	.503	480	104.0
Thorndike Scores		89	12.96

COMPARISON OF PREDICTIVE VALUES OF NEW AND OLD PLANS OF ADMISSION

In judging the worth of a new instrument it is often more meaningful to compare its results with those of the old instruments which it is to displace or supplement than to learn how far short of perfection its results fall. The presentation of the data which follow is not prompted by any desire to make invidious comparisons, but by the desire to define more clearly the efficiency of the intelligence test as a criterion for admission to college.

In comparing the predictive value of the Thorndike Examination with that of the traditional content examination, it is necessary to describe the particular content examination used in order to arrive at any intelligent

appraisal. The comparisons here made are with New York State Regents' Examination marks and high school marks.

The Napoleonic form of educational organization which distinguishes New York State from all other states in this country gives the Regents' examinations great advantages over any other system of content examining known to the writer. The Regents' examinations are prepared by a central office on the basis of a minutely defined curriculum which is followed by all State Schools under the supervision of an efficient corps of inspectors. These examinations, therefore, give comparable measures of achievement in material to which all examinees have been approximately equally exposed, and are therefore quite efficient and reliable indicators of differences in native capacity.

TABLE 24

Correlations between Thorndike scores and scholarship scores earned in indicated periods of time by 233 men admitted under both Old and New Plans in September, 1919, who remained in Columbia College two full years.

	r	M	sigma
Freshman Year	0.562	230	56.2
Sophomore Year	.500	230	66.0
Freshman and Sophomore Years	.592	460	112.0
Thorndike Scores		87.4	12.84

The marks from high schools outside New York State, on the other hand, suffer not only by a want of any common standard, but also by a lack of agreement as to the basis of marking. Not only are examinations of the same name variable, but the content included in courses of

the same name varies from one school to another. A score of sixty in one school may be equivalent to eighty in another because of differences in the conventional interpretation of such percentages, in addition to the other

TABLE 25

Number of men	Regents' Examination Scores	Scholarship Score	Quality Index for 64 Hours	Number of Hours of College Work in Terms of Each of the Three Passing Grades		
(1)	(2)	(3)	(4)	(5) C	(6) B	(7) A
2	50-69	340	5.31	57	43	31
50	70-79	405	6.33	67	51	37
64	80-89	455	7.10	76	57	41
28	90-100	615	9.61	105	79	57

TABLE 25. Quantity and quality of college work done by 144 Old and New Plan men admitted in September, 1919, who had taken the Regents' examinations of the State of New York. This Table is similar to Table 12, q.v. It is to be read as follows:

There were two men admitted with Regents' scores between 50 and 69; the average S.S. of these two men for two years in Columbia is 340; on the basis of a normal two-year program of 64 Hours, their average quality index is 5.31, or C—; in terms of C grade work, they completed in two full years of residence, 57 Hours each; in terms of B grade work, they each completed 43 Hours; in terms of A grade work, 31 Hours. The 28 men with Regents' scores between 90 and 100 who were admitted in September, 1919, and who remained in Columbia two years, have an average S.S. of 615; the average quality index on the basis of a normal program of 64 Hours is 9.61, or A—; in terms of C grade work, each of these 28 men completed 105 Hours of credit work; in terms of B grade work, 79 Hours; and in terms of A grade work, 57 Hours, which is just the number completed by the 50-69 men at C grade.

large unreliability factors in the traditional content examinations.

That the Regents' examinations are good criteria for admission to college is apparent in Tables 25 and 26; while the inferiority of high school marks is patent in Table 27.

The predictive value of the Regents' examinations is almost identical with that of the Thorndike intelligence test, as may be seen by comparing Tables 12 and 25.

TABLE 26

Regents' Scores	Percentage of Men Making 7 or More F or D Hours in Two Years
50- 59	100
60- 69	100
70- 79	49
80- 89	29
90-100	10.7

TABLE 26. Percentages of men at various levels of Regents' scores who in two years receive a grade of F or D in 7 or more Hours of college work, based on the records of 144 Old and New Plan men admitted in September, 1919, who remained in Columbia two years.

TABLE 27

High School Marks	Percentage of Men Making F or D in 7 or More Hours in Two Years
60- 69	40
70- 79	50
80- 89	42
90-100	24

TABLE 27. Percentages of men at various levels of high school marks who in two years receive a grade of F or D in 7 or more Hours of college work, based on records of 103 Old and New Plan men admitted in September, 1919, who remained in Columbia two years and whose secondary school records were available.

Table 26 shows the same high predictive value of Regents' examinations in a reciprocal way. But it is only fair to recall that the time for a complete Regents' examination averages at least fifteen hours, while that for the Thorn-

dike Examination is now not much above three and will hereafter be less than three.

The poverty of the high school marks in predictive value suggests at once the urgent need for the adoption of a common or comparable system of units and measurements by secondary schools. The measurements made by a large number of college entrance examination agencies are more or less vain attempts to do that which the secondary schools are supposed to have done and which they should have done. They spend enough time, effort, and money

TABLE 28

Correlations between the number of Hours of F or D grade work done in two years in Columbia and scores in each of three different criteria for admission to college, based on records of men admitted in September, 1919, who remained two years.

	r	N
Thorndike Intelligence Scores	—.465	233
Regents' Examinations	—.432	144
Secondary School Marks	—.205	103

to do it adequately, but nearly all of it is wasted in futile individualism. The adoption of a uniform system of units would in no wise impair local autonomy nor destroy local initiative. All the high schools in the country should report their grades in units as comparable as are those used by the high schools within New York State. The difficulties appear insuperable to many leaders now, but it is a problem worthy the best thought and intelligence in our school administration.

The percentage of various intelligence and Regents' groups that secure F or D grades in a given number of Hours of college work is a simple and striking way of

indicating predictive value, but the technical reader may demand a more exact measurement of the relation between "number of hours failed" and scores on the various college entrance criteria.

Table 28 presents evidence of this relationship in terms of Pearson correlation coefficients. Since the correlations are between number of points failed and intelligence scores, they are negative, and the greater the negative correlation the greater is the predictive value, because a large negative correlation means that men with high scores get relatively small numbers of F and D grades. The most

TABLE 29

Correlations between two-year scholarship scores in Columbia College and scores in three different criteria for admission.

	r	N
Thorndike Intelligence Scores	.672	111
Regents' Examinations	.644	144
Secondary School Marks	.262	103

favorable correlation is given by the Thorndike Examination, with Regents' a close second and high school marks very low. The comparative predictive values of these three criteria are more convincingly set forth in a positive way in Table 29.

Table 29 is based on records of 233 Old and New Plan men admitted in September, 1919, who remained in Columbia two years; but the records of all these 233 men are not represented in each correlation. The correlation for the intelligence examination involves only the men admitted under the New Plan; the correlation for Regents' involves all men who had taken Regents' examinations,

whether admitted on the Old or New Plan; and the correlation for high school marks involves all men whose school records were available. Again the three-hour intelligence examination is seen to be as good as the fifteen-hour content examination, while the high school record is

TABLE 30

Correlations between first-year S.S. and score in four different admission criteria.

	r	N
Thorndike Intelligence Scores	.59	199
Regents' Examinations	.57	34
College Entrance Examination Board	.432	35
Secondary School Marks	.331	67

TABLE 31

Correlations between S.S. of Sophomore fall semester and score on three different criteria for admission.

	r	N
Thorndike Intelligence Scores	.45	283
Regents' Examinations	.40	153
Secondary School Marks	.15	141

very inferior as a criterion for admission to college. These relationships are confirmed by two other measures shown in Tables 30 and 31.

A COMBINATION CRITERION FOR ADMISSION

It has been said in preceding chapters that the intelligence test would find its greatest usefulness when used in

conjunction with all other data available about candidates and students. Since we have several criteria, each of which has considerable predictive value alone, the question arises as to what is the best possible combination of them for purposes of general admission to college. This question cannot be answered reliably here because only ninety-seven cases could be found in which the records included indices of the three kinds desired, namely, intelligence scores, Regents' scores, and high school marks. In

TABLE 32

	I	II	III	IV
I. Freshman Year S.S.	1.00	.55	.55	.29
II. Thorndike Intelligence Score	.55	1.00	.40	.12
III. Regents' Examination Score	.55	.40	1.00	.424
IV. Secondary School Mark	.29	.12	.424	1.00
Proportional Weights		1.00	1.00	.2269
Multiple correlation after each variable is added		.55	.6573	.6623

TABLE 32. Determinant formed by the intercorrelations between four indicated variables, showing proportional weights and total correlations as solved by Toops' formula. Based upon records of 97 men who remained in Columbia College at least one year.

order to get as many as ninety-seven cases, it was necessary to limit the scholarship score to one year. Intercorrelations of these four variables for these ninety-seven cases were calculated with results as in Table 32.

Avoiding intricate mathematical discussions, it is enough for the general reader to whom all the implications of this table may not be clear, to learn that the proportional weights which the Thorndike and Regents' examinations should have are equal, while that of high school marks is more than one-fifth but less than one-fourth of the weight which either Thorndike or Regents'

scores should have. The combination of Thorndike and Regents' scores, equally weighted, raises the total correlation from 0.55 to 0.6573, while the addition of high school scores to the combination with the best possible weighting raises the total correlation further to 0.6623,—an increase of only 0.005. It is unnecessary to add that the weighting earned by the Regents' Examinations is almost certainly higher than the weighting that would be earned by any other content examination system now given for college entrance purposes in the country. The reason for this has already been mentioned: the Regents' Examinations are given only to those for whom they have been specifically designed, and only to those who have had opportunities for specific preparation as nearly equal as could be achieved by a highly centralized educational organization.

The combination of Thorndike and Regents' scores has been made with approximately equal weighting for 144 cases of men who remained in college two years. The correlation of this combination with two-year scholarship score is above 0.65, thus confirming the theoretical results obtained in Table 32.

OTHER COMPARISONS

Comparisons might also be made as to the predictive values of other traditional criteria, such as estimates based on personal interviews, Principals' recommendations, employment history, health records, etc. But such measures were not available for this study.

It is to be regretted especially that no measures of "personal qualities" were available for this study. In the case of Columbia College candidates, the personal interview with the Director of Admissions is a large factor.

An attempt was made to keep records of the personal estimates of the Director of Admissions on the members of the entering class of 1920; but various factors prevented the achievement of any satisfactory results.

The character records supplied by the high schools, and the letters of personal recommendation submitted by the candidates proved to be fruitless also. The meaning and strength of adjectives, their moralistic tone and coloring, seem to vary greatly from one locality to another. The

TABLE 33

Correlations between Thorndike scores and indicated Morphological Indices.

	r	N
Length of limbs	.155	75
Volume of trunk	—.36	75
Morphologic Index (L.L. divided by V.T.)	.356	75
Height	.0415	225
Weight	—.1830	225
Ratio height to weight	.2300	75
Lung Capacity	—.1830	225

writer derived the curious impression (from a somewhat inexact examination, be it admitted) that those boys whose recommendations were most profusely adorned with glowing adjectives and subtle compliments were most likely to be frequent callers at the Dean's Office around examination times. One such case became notable because of the synchronous relations maintained over a long scholastic period between the incidence of "severe" but "temporary" illnesses and the scheduled examination periods of the college.

It was originally planned to study the relation between health and scholarship, but it has not been possible to give

any effective attention to this very important question. A most interesting and suggestive study of a question not unrelated to health has been made by Dr. Santé Naccarati, and published in the Columbia University Contributions to Philosophy and Psychology, No. 45, August, 1921, Volume 27, No. 2, under the title, "The Morphologic Aspect of Intelligence." The writer had the pleasure of cooperating with this young Italian scientist to the extent of furnishing several sets of Thorndike intelligence scores,

TABLE 34

Correlations between indicated groups of tests in the Thorndike Examination and Morphologic Index and Ratio of height to weight.

	Morph. Index	N	Ratio: H/W	N
Mathematical	0.275	67	0.256	192
Reading	.251	67	.246	192
Completion	—.023	67	.036	192
Logic (syllogisms)	.249	67	.252	192
Non-Verbal	.055	67	.104	192
Part I (D plus E)	.252	67	.235	192

and of reviewing some of Dr. Naccarati's statistical findings. Some of the relationships found by him, and confirmed by myself, are presented in Table 33, with his permission.

The highest correlations here are quite significant. It was thought that morphologic measurements might have special relations with special mental abilities, and correlations were secured between various groups of Thorndike tests and two morphological factors with results as in Table 34.

The results are interesting, but not convincing enough to form the basis of any significant remarks. Our main

interest as college administrators is not the relation between morphology and intelligence as measured by any particular test, but between morphology and success in college. The writer is indebted to Dr. Naccarati for the morphological data involved in Table 35, which shows an almost total absence of relationship between morphology and college success.

TABLE 35

Correlations between First Semester Scholarship Scores and indicated morphological indices.

	r	N
Morphologic Index (Ratio Length of limbs to Volume of trunk)	.123	67
Length of limbs	.080	67
Volume of trunk	—.090	67

CONCLUSION

The intelligence test is not only as good a criterion for admission to college as any other single criterion thus far used, but it is more efficient and less expensive. Its nearest rival is the leaving examination of the New York State High Schools. The strength of this leaving examination lies in the fact that it approximates an intelligence test, in addition to measuring actual achievement, industry, etc.

CHAPTER FIVE

Analysis of the Thorndike Examination

The analytic work on the Thorndike Examination has been handicapped seriously by two facts; first, the unreliability of the scores on the individual tests, due largely to the variable amount of time which in practice is actually spent upon each test; and second, the unreliability of grades in college courses, due mainly to subjectivity in all its ramifications—variable standards of achievement, subjective scoring, indefiniteness and variability of the bases of grading, inadequate samplings of materials and performances, the inclusion of irrelevant and absurd considerations, etc. The results secured thus far are therefore largely negative.

The possible uses of analysis appear to be quite numerous. In this chapter reference will be made only to a few of the most obvious possible uses:

1. The deriving of "secondary checks" for use within the "critical area" for admission to a college in which registration is strictly limited.
2. The deriving of differential admission criteria for the various departments of a modern university.
3. The deriving of standards for registration in individual college courses, e.g., English and Mathematics; such standards so designed as to differentiate between those who have a special disability and those who are merely deficient in preparation.
4. The deriving of standards for adjusting the academic load of each student in view of his general intelligence, his special abilities and disabilities, his preparation, and the amount and character of his necessary concurrent outside activities.
5. Vocational guidance in the broadest sense.

6. Improvement of the Thorndike Examination by the elimination, revision, or re-weighting of individual tests on the basis of their correlations with the criterion.
7. Research.

The possibility of developing "secondary checks" within the intelligence examination depends upon the presence of individual tests or groups of tests that have special differentiating powers within the "critical area" of the total examination scores. The desirability of securing such checks is attested by a study made with that end in view by Mr. Harold K. Chadwick, Assistant Director of Columbia University Admissions. The object of the study was "To find an objective check to aid in selecting those men to be admitted from the group receiving Thorndike Examination Total Scores between fifty and seventy, the check to be based upon a study of those parts of the intelligence examination proved to be critical in this connection." [1]

The character of the college work accomplished in the second semester by the men admitted with intelligence scores between fifty and seventy who had survived the first semester is shown by the fact that twenty-one of the sixty-eight men in this category failed or withdrew, fourteen secured an average grade of D, seventeen secured an average grade of C, and seventeen B or better. One-half the group did "satisfactory work," and one-fourth did good work. To quote Mr. Chadwick further:

> Probably a third of these men were worth having in college. The group probably represents a selection from about 150 men. While those excluded would not furnish the same proportion of desirable men, a secondary check would certainly have been of use in making a proper selection. As the total number of really doubtful cases did not exceed 210, the check

1 From an unpublished study by Mr. Chadwick.

would probably affect two-thirds of the cases requiring careful decision during one registration period.

The college records of thirty-six of sixty-eight men were studied carefully; twenty should not have been excluded, and the other sixteen included only two men whose records were above question. The records of these two groups of twenty and sixteen men on the intelligence examination were then carefully studied. Their scores on Tests 5, 7, and 12 of Part I gave no decisive results. Tests 1 (Reading) and 8 (Information) of Part II, taken in conjunction with the other three mentioned, gave somewhat better results. If a man failed to reach the arbitrary standard set up for each of these tests, his exclusion would have been justified; but very few of the men failed to reach the arbitrary standard in all the five tests chosen for this purpose. The conclusion is that the use of these five tests as secondary checks would have resulted in slightly better results on the whole. But in order to make these checks valid, the critical scores on each of the tests thus specially used must be fixed by a more rigid method of investigation than has been used in this brief preliminary study.

This highly suggestive study suffers by several inadequacies, all of which Mr. Chadwick clearly recognizes and warns us against. In the first place, sufficient care is not taken of negative instances in the derivation of his checks. It is true that the check desired is only for those who receive total scores between fifty and seventy, but the derivation of the checks need not be based solely upon the records of men in this category. If the individual tests have any dichotomous propensities, they could be securely defined only by testing their operation in all ranges of general intelligence. It was with this end in view that the differentiating powers of various tests and groups of tests were studied with unselected samplings of men admitted to Columbia College.

Table 36 shows distributions, means, sigmas, and correlations with two-year college scholarship scores, of eight individual tests in the Thorndike Examination, namely,

TABLE 36

Score	Tests in Part I					Tests in Part II		
	1	2	6	7	13	3	5	7
−6						1		3
−5								
−4								3
−3								
−2						1		9
−1								
−0						1		
0		1	1			1		12
1	2							
2	4					5	1	8
3	10			9				
4	29	2	1	34		4		10
5	66			68				
6		5	2			7	1	6
7								
8		14	5		1	9	2	10
9		3						
10		86	6		3	12	3	18
11								
12			4		7	21	9	13
13								
14			18		12	15	12	3
15								
16			17		12	20	22	4
17								
18			38		38	13	39	4
19								
20			19		38	1	22	3
21								
22								4
23								
24								1
N	111	111	111	111	111	111	111	111
M	4.2	9.2	15.8	5 0	17.4	11.6	16.4	7.0
sigma	1.9	3.1	3.9	1.3	2.9	5.3	3 3	7.7
r with 2-year S.S.	.253	—.055	.36	.08	.17	—.14	.38	.13

TABLE 36. The five tests of Part I are: (1) Written Directions; (2) Disarranged Sentences; (6) Opposites; (7) Judgment, or "Best Reason"; and (13) Matching forms or form recognition. The three tests of Part II are: (3) Pictorial Completion; (5) Geometric Form Series; and (7) Trade Questions.

Tests 1, 2, 6, 7, and 13 of Part I and Tests 3, 5, and 7 of Part II. One of the tests which Mr. Chadwick chose was Test 7 of Part I. Of 111 men who survived at least two years in college none received less than three points on Test 7; and its correlation with college success is practically zero. It is possible that a test with little differentiating power in the whole range of college talent may have great efficiency in the lower ranges; but it seems quite improbable that this particular test, even on a strictly uniform time allowance, could be trusted very deeply as a basis for differentiating reliably in any range. With only five very easy chances to win or lose, it is quite probable that the failure to get a perfect score can nearly always be explained as the result of pure chance; certainly an appreciable number of such failures *is* due to chance, and the bright boy anxious to make speed and made confident by the great ease of the problems of Test 7 is just as likely to make chance mistakes as the stupid plodder, if not more likely.

The correlations for Tests 2, 7, and 13 of Part I, and 3 and 7 of Part II are so nearly zero as to condemn these tests in their present forms as unworthy of the time allotted to them in the total examination. The tests of "scrambled" sentences and pictorial completions actually give negative correlations, indicating probably that the bright boys pass them over as bad employment and that the stupid boys waste much time on them for fear of flying to other ills of which these may be only a mild foretaste. Test 10 (Number checking) of Part I and Test 4 (Pictorial analogies) of Part II also give correlations with college success so near zero as to afford no reliable differentiation in any range of college talent. Apparently no one of these tests holds out any promise of affording special differentiation in the critical area.

TABLE 37

	Thorndike Test Groups:							Total Thorndike Score	Regents' Marks in indicated subject matter
	Language	Mathematics	Information	Reading	Trade	Completion	Logic		
	1	2	3	4	5	6	7	8	9
Two-Year College Scholarship Score	.40	.57	.48	.51	.19	.35	.42	.67	
Two-Year Average College Grades in:									
English	.42	.24	.35	.39	—.15	.46	.18	.47	.27
French	.24	.65	.45	.32	.19	.22	.47	.65	.40
Chemistry	.22	.38	.17	.19	.08	.11	.07	.40	
Physics	.27	.30	.28	.33	—.15	.34	.35	.43	
Mathematics	.35	.66	.38	.37	.20	.31	.39	.68	.47
Contemporary Civilization	.42	.27	.26	.42	.12	.46	.28	.56	

TABLE 37. Correlations between indicated variables for groups of students varying in number from 70 to 144 from the 233 Old and New Plan men admitted to Columbia College in September, 1919, who remained in college two years or more. For each correlation all the men having scores in both indicated variables were used. N is always above 100 except for the Chemistry and Physics correlations. Contemporary Civilization is a one-year course, and only in the cases of repeaters are the grades based on more than one year's work. In the other college courses most of the grades are based upon two full years of work.

In order to learn the relations between groups of tests and college success (general and analyzed), the correlations of Table 37 were calculated.

The interpretation of these data must take account of the reliabilities of the variables involved. Our measures of the reliability of the grades in college courses are not wholly satisfactory, since they are indirect. Our best available measures are the correlations between two semester's grades in each course. They are all probably too low, but are given here for what they are worth:

TABLE 38

Correlations between first and second semester college grades in indicated subject matters.

	r	N
English	0.36	170
Contemporary Civilization	.57	140
Mathematics	.61	85
French	.68	140
Chemistry	.39	25
Physics	.40	35

The reliabilities of the groups of Thorndike tests were given in Chapter Three:

Language	0.68
Mathematics	.81
Reading	.48
Completion	.57
Total Thorndike Examination	.85

With these coefficients in mind we may proceed to a more intelligent appraisal of the data of Table 37.

Looking at the first row of correlations, we find that the Mathematics Group of Thorndike tests gives a correlation of 0.57 with general college success. When we consider that this group requires only twenty-six minutes

to give, its power to predict success in college two years ahead is truly remarkable. The Reading and Information tests are next in order of prophetic powers. The Reading test apparently correlates with college success better than with itself, but this is explained by the fact that its reliability as here given is the correlation between the two halves; the self-correlation of the whole would undoubtedly be greater than 0.51, though not as great as would appear by the application of Brown's formula. The Trade test of Part II is the lowest in this group in power of predicting general college success.

Three different groups of Thorndike tests predict average marks of two years of college English better than marks for the first session of college English predict marks in the second session. The correlation between first and second session English marks is 0.36; that between the Regents' English Examination and two-year college marks in English is 0.27. The correlation of two-year college marks in English with scores on the sixteen-minute Language Group in the Thorndike Examination is 0.42, and with scores on the thirty-minute completion test it is 0.46. In view of the great emphasis put upon "passing" the first session of college English, these correlations are very significant.

A very curious tradition which is partly self-contradictory survives in the English departments of most of our colleges. English departments everywhere explain the low correlations between first and second session English marks as being true pictures of things as they are and should be; that is, they put forth the doctrine that success in the first semester does not imply any great probability of success the next semester in English, that many courses *called* English really represent quite different activities and subject matters. This curious doctrine is put forth in the face of the very rigidly enforced rule that no one may even register for second session English

who has not creditably "passed" first session English. In so far as this rule implies that success in first session English is prerequisite to success in second session English, it contradicts the doctrine that is brought forth to explain away the low relationship between marks in English gained in successive sessions.

The fact of the matter seems to be that there is really a very *high* relationship between English performances in two or more sessions, and that the explanation of low correlations is to be found almost wholly in the unreliability of traditional marks in English courses. This fact cannot justly be taken as a reflection on English teachers. They recognize the weaknesses of English grades from the viewpoint of accurate and scientific measurement no less than their critics. They know that the marks are thoroughly subjective, that there are large and frequent variations in standards, that the elements that enter into grades are frequently irrelevant and always variable; and they are as anxious as any others to improve their grades. But, although some promising suggestions have been offered, no workable program for the accurate measurement of the particular values claimed for English courses, such as exposition, style, literary appreciation, etc., has anywhere been presented for the judgment of competent thinkers. It is true that certain essential elements of these values can be measured accurately, such as the informational background upon which any useful application of the higher values must depend. Some unthinking opponents of new methods of measurement have gone so far as to say that they care not a whit for "mere information"; but no intelligent teacher of English will be bothered seriously by so absurd a generalization. After all, we cannot appreciate literature or any other facts of experience without having those facts in mind. If ten novels have been assigned to a class, the mere act of reading those novels in a factual way has some value; and practically the only means of ascertaining objectively and reliably whether all of the ten novels have been read is the new and efficient form of examination illustrated in succeeding chapters.

Just how and to what extent these new forms of examinations can be used in English courses is a question which must be resolved by English teachers. The experts in measurement may be of some assistance in bringing into play the necessary

experience gained in other subject matters; but it will primarily be a task for English teachers in which caution must prevail over haste if any good results are to be attained.

Referring again to Table 37, we find indicated in the third row of correlations, specialized predictive powers for French in a higher degree than we found in the row of correlations for English. The correlation of grades on the Regents' French Examination with college marks in French is 0.40. The twenty-six-minute Mathematics group of Thorndike tests correlates with the average college mark in French for two years to the extent of 0.65. This astonishing relationship was checked up three times with different groups of students, so that there is small chance of error. The population for this coefficient is ninety-three, which included all the Freshmen who came into college in 1919 and had had two years of French in college by June, 1921. The Thorndike Logic and Information Tests give correlations with two-year college marks in French of 0.47 and 0.45. These high correlations are due at least in part to the high reliability of college marks in French. The French department has been using some of the new forms of questions for several sessions in the non-literary courses in which the main emphasis is on learning French as a medium of communication. As a result the reliability of French marks, as measured by the correlations between grades for two sessions, is not less than 0.68. A more complete exploitation of the new forms of examinations will probably raise the reliability of the marks to 0.80.

In the fourth row of correlations in Table 37, we find the coefficients lower, on the whole, than in the second row. This, again, is undoubtedly due in large measure to the unreliability of the chemistry marks. The correlation between the marks for two semesters in chemistry is

0.39, and the highest correlation that chemistry marks give with any of the Thorndike groups is that with the mathematics group, 0.38. This is too low to afford any secure guidance, and is an excellent example of the need for developing a more reliable marking system.

The row of correlations for physics presents the same story as that for chemistry, and needs no further comment.

The high correlations for the two-year college grades in mathematics are associated with high reliability of college marks. The correlation between first and second session college grades in mathematics is 0.61; the reliability of two-year marks in mathematics is at least 0.70. The correlation of Regents' Examination marks in mathematics with two-year college marks in mathematics is 0.47. The correlation of two-year college marks in mathematics with scores on the twenty-six-minute group of mathematics tests in the Thorndike Examination is 0.66. That is, a twenty-six-minute test predicts success in college mathematics better than the first session college mark predicts the second session college mark in mathematics, and much better than the Regents' Examination in mathematics predicts success in college mathematics. This is almost convincing evidence of the inadequacy of the traditional content examination in mathematics. The issue is squarely drawn on experimental grounds; there is no means left to mere subjective opinion and conservatism whereby these may claim for the old examination in mathematics merits which it does not possess. There is no longer any excuse for tolerating a mathematics examination which predicts college success in mathematics to an extent no greater than 0.47 when we have examinations which give a prediction as high as 0.66.

The Completion, Reading, and Language groups give

the highest correlations with Contemporary Civilization marks. The reliability of the Contemporary Civilization marks used in the correlations of Table 37 is 0.57. These marks were derived under the old system of marking. The reliability of Contemporary Civilization marks under the new system of examining in Columbia College described in the following chapters, is about 0.90, and their correlation with intelligence scores about 0.69. As was to be expected, no one test or group of tests has very special predictive power for success in Contemporary Civilization; it is a very general descriptive course, involving a great deal of critical reading. It is no surprise, then, to find that the linguistic group of tests correlates somewhat more highly with Contemporary Civilization marks than other groups.

The general conclusion suggested by Table 37 is that admission (or perhaps registration) should be differential, not general. The correlations of groups of Thorndike tests with grades in individual subject matter courses show plainly the possibility of advising students profitably at registration on the basis of their standing in these groups of tests. Certainly the correlations of 0.65 and 0.66 with French and mathematics grades in college would give warrant for rather definite advice or requirements from students entering those courses. And it is probable that equally strong bases for advice with reference to other college courses are concealed by the low reliabilities of the college grades in those courses.

These conclusions are only strengthened by the correlations in the last column of Table 37. It seems evident that the correlations of the Intelligence test are limited by the reliability coefficients of the grades in the several college courses. Thus we have the curious situation of getting a higher correlation between general intelligence and mathe-

matics grades than between general intelligence and the two-year college scholarship score.

For a group of 106 New and Old Plan men admitted in 1919 who remained in Columbia College two years, the correlations of Total Thorndike Intelligence Scores with the following college indices are as follows:

Two-year scholarship score	0.594
Two-year average of mathematics and French grades	.599
Two-year average of mathematics and Con. Civ. grades	.658

The adjustment of academic load on the basis of intelligence scores is justified by the high relationship between such scores and reliable measures of college work. We know, for example, that the men with high intelligence examination scores actually do more college work in a session than the men with low scores. Moreover, we know that the men who accomplish less than the brighter students are unable to do more; they do not voluntarily abstain from getting credit for high achievement. Yet we allow, if we do not actually encourage, all men to attempt about the same amount of work per session. That much lost motion, fruitless or wasted effort, and expense occur is evident from these correlations for 103 pre-Medical and pre-Engineering Freshmen:

Thorndike Intelligence Scores with points earned	0.540
Thorndike Intelligence Scores with points attempted	.120
Thorndike Intelligence Scores with Hours of outside work, commuting, etc.	.000
Points attempted with outside work hours	.003

On the theory that men would not only be happier but would do more and better work if they were required to attempt only as much as they had a reasonable chance of coping with, the Faculty Adviser of the pre-Law, pre-Journalism, and pre-Business students attempted to adjust their academic undertakings somewhat in accordance

with their intelligence scores and outside distractions. Two results were immediately observed.

The first was that the correlation of the amount of college work attempted with intelligence scores rose to 0.35, and with amount of time regularly spent on outside activities (work and commuting) to — 0.25. The negative correlation indicates that the men with many outside distractions attempted less college work, and that the men who had more free time attempted more college work than the average. This adjustment was accomplished without materially changing the average amount of work attempted, and only with a slight increase in variability. The Means and Standard Deviations for the two groups are as follows (the units are "Points" or "Hours" of college work):

	M.	S.D.
Pre-Medical and Pre-Engineering	18	1.54
Pre-Law, Pre-Business, and Pre-Journalism	17	1.64

An important detail to notice here is that as many were advised to take heavy programs as were advised to take light programs. That is, the object was not merely to keep the less able students out of deep water, but also to keep the bright boys out of idle mischief.

The second result cannot be objectively measured, but is one which is no less significant than obvious to the adviser of the students concerned. The less able boys have fewer failures to discourage them, do better work in what they do attempt, and are far happier; while the brighter boys who were prevailed upon to take heavy programs have the satisfaction of accomplishing more than they thought possible. The indirect effects of keeping these bright boys busy are more important in the long run than lessening the number of failures in the lower ranks. This modest experiment suggests that a small but very

important problem in collegiate educational guidance may be partly solved by the aid of the intelligence test. At the present time students may register for any number of points of college work between twelve and nineteen, inclusive, without restriction. The average number of points attempted by Freshmen is 17.5, and the average number "passed" is about 15. Within the range of twelve to nineteen, the number of points to be attempted by each student might be left to the student's adviser, who would take into account all relevant factors, such as intelligence examination score, preparation, character, time to be spent in making a living, commuting, etc. Many of the most pathetic failures in college that have come to the attention of the writer have been cases of boys with average or higher ability who were allowed to take a normal program when they were compelled to work for a living forty hours per week, and to spend from two to four hours on crowded commuters' trains each day. The number of such cases could certainly be reduced by a cautious use of the intelligence examination scores in conjunction with the other relevant factors mentioned above.

The analysis presented in this chapter throws some light on the problem of improving the Thorndike Examination as a collegiate administrative instrument, both in the way of increasing its significance and reliability as measured by its correlation with the best criteria available, and in the way of decreasing its cost in time and money. Concrete suggestions growing out of this analysis and supported by the data in all the preceding chapters are presented in Appendix III, pp. 320 to 330.

CHAPTER SIX

College Success

WHAT IS COLLEGE SUCCESS?

The apparently simple and glibly used words "college success" are in reality the symbols of a large group of ideas which are complex individually and collectively and quite variable from place to place and from time to time. In the last chapter we rather freely identified our Scholarship Score with college success; but this assumption is justifiable only if the limitations of the identity are understood and kept in mind.

College life is a very complex and variable phenomenon; in the strict sense, it is as broad and imponderable as life in general, and success or failure in college can be defined only as life in general can be defined. College life is a matter which can be discussed satisfactorily only from the viewpoint of philosophy in its largest and broadest sense. This fact, so keenly appreciated by experts in mental and educational measurement, does not mean that the specialist has no place in educational administration;—it does not mean that quantitative science is barred out. After all, we cannot escape the fact that whatever college success is or is not, it follows the general rule that whatever exists at all exists in some quantity or amount. The effort to measure the quantity of a phenomenon is but an extension of the effort to learn what it is. It is no more unphilosophical or undignified to subject education to quantitative treatment than it is to subject interstellar space or force to quantitative study. Some philosophers and others denounce the psychologist and statistician as one who tries to *measure* purely *qualitative* phenomena such

as *culture* and *education.* The position of such objectors with reference to measurement in education, in so far as it has any logical status, involves the assumption that culture and education are mysterious elements whose values are dependent upon their mystery and indefiniteness. The effort to measure a complex phenomenon is not an effort to rob it of any value that may inhere in its complexity, but rather an effort to understand the phenomenon for the purpose of increasing human control over it. Certain factors in a complex may be measurable, though all are not. The fact that some are not measurable should not keep us from measuring those that can be measured.

While college experience is very wide and variable, and is part and parcel of life in its most inclusive sense, being valuable mainly for that very reason, the college nevertheless has certain special characteristic functions which distinguish it from other institutions, the performance of which may be said to measure its success as a distinctive institution. An inventory of these functions might be presented in detail, but it is enough to mention four classes into which they fall in order to convince the reader that the expert in measurement is not overlooking any essential element of college success:

(1) Of recent years the physical competitions have been greatly emphasized in the college curriculum. This movement has received a great impetus alike from the new and increasing interest in public health among intellectual people, and from the growing realization of the spiritual values of athletic contests. The advertising possibilities inherent in athletic renown have been exploited to the limit by institutions everywhere; but it would be absurd to attempt to explain college athletics as a result of sordid competition in advertising. The athletic field of the modern college is rather to be explained as a happy re-

crudescence of the great national games of Ancient Hellas, whose culture we are proud to emulate. No one who has sat in the Yale Bowl can doubt the supreme spiritual stimulus of those chivalrous contests in which the losers no less than the winners are crowned with glory. The less spectacular congregations in smaller and remoter centers where elevens are pitted against elevens do not suffer as much as is commonly supposed by comparison with the Yale Bowl in respect to noble stimulations. The games suited to smaller audiences, and actively participated in by larger numbers of athletes, have the same spiritual significance on a somewhat smaller scale.

The stimulus of these games for better health is as vast as it is undoubted. Even here the higher aspirations of our culture are served directly. No one can estimate the indebtedness of Phidias or Praxiteles to the ancient institution of Greek national games in the production of those matchless models that have been the inspiration of artists through the ages since the days of Hellenic supremacy; but no one doubts the largeness of that debt.

Participation in this phase of college life, whether on the gridiron or in the bleachers, is one of the finest elements of collegiate experience. To ignore it in estimating "college success" is to leave out of account one of the most noble and potent achievements of the college. The more formal work in physical education not only gives needed information about personal hygiene in general, but develops hygienic habits of eating, sleeping, exercising, and of cleanliness, which in their ultimate results rise far above the purely physical well-being of the students in college and in post-collegiate life.

(2) The development and encouragement of conduct, habits and attitudes that are socially desirable, both from the viewpoint of immediate social intercourse and from

that of remote social goals, is a prominent function of the college. No one who has witnessed the transformation of angular and awkward Freshmen into smooth and well-poised members of a social community in which they have learned to get along with their fellows, can doubt the efficiency of even our smallest colleges in this respect. The contacts made in the office of admissions, in the registration room, in the class room, in the libraries, in the dormitories, in the assemblies, in the "activities," whether on the athletic field, in the bleachers or grandstand, in the gymnasium, in the college daily editorial office, in the dramatic societies, or in the college halls and elevators,—these are the destroyers of provincialism, the curbers of unrestrained individualistic instincts, the creators of social habits and attitudes that enable their possessors to get along with other people in the best and most economical way. It is these that act as balancers, and rescue the single-track book-worm from pedantry and snobbery. The obsessions of the book-worm may be just as hateful and selfish as those of the unlettered boor from the hill-country; and for such the "extra-curricular" activities are probably more soul-saving than the prescribed courses of study.

(3) Professional preparation for leadership in the work of the world is an end which our liberal colleges of today are unashamed to serve. This preparation also consists of "prescribed" and of "extra-curricular" activities and contacts. Professional and technical information and practice are taught in class work, and the personal and human elements are developed in the general socializing life of the college.

(4) What might be called the most important and most characteristic function of the college is the introducing of students into what President Butler has happily called

"The Great Tradition" (Address at opening of 168th session of Columbia University). In its broadest sense this includes our heritage of accumulated human experience in its totality in every branch of human interest, but more especially it refers to the cultural and liberal arts and sciences that particularly characterize the interests of western civilization.

No clear line of division may be drawn between these four functions; they are all aspects of the same thing, and the methods of attaining each of the goals consist so largely of common elements that no hard and fast divisions are possible. But no one will question the predominance of academic and intellectual factors in the practical achievement of professional and cultural education. This is the *raison d'etre par excellence* of the college, and the main line of attack is the academic one. Since this is so, we may lay down the general principle that the most important measure of college work is its academic product.

This does not mean that other products, such as come from the gymnastic and social activities of the college, are unimportant. But it seems perfectly clear that the specially characteristic task of the college, the one which particularly distinguishes it from other social institutions, is the academic task of acquainting the rising generation with the civilization, with The Great Tradition, which is our heritage and theirs. Athletic and social activities, while necessary and indispensable phases of college life, are not its special purpose. The country club and the Tuesday club might serve these purposes as well as or better than the average college.

The *sine qua non* of the college, then, is its academic work. It is for this work that learned men and experts in every division of science and art are collected from every part of the world. They teach us the traditions and

stored achievements of the past, and they teach us how to carry on the noble work of adding to this vast and venerable store of human learning, achievement, and culture. Learning and Research—these are the yardsticks which measure the highest service of the college. There is no thought anywhere of ignoring the other services of the college; but it is maintained that the best and most reliable measure of college work is its academic products.

CHARACTERISTICS OF COLLEGE MARKS

It has long been the custom of educational institutions to rate students nominally in the order of their academic achievement. But the definition of this achievement has been very vague and variable.

College marks have presumed to measure many things which could not and cannot yet be measured with any degree of accuracy. The spiritual factor, which forms so large a part of what the college strives after, is an imponderable one for which no intelligible definition, and consequently no adequate scale, is available. Yet many instructors now as in the past hesitate not to pass judgment upon what is in the heart of man, knowledge of which, according to the Bible, is reserved to God alone.

The efforts of colleges to be quantitative have led many into paths of heroism which the angels traditionally avoid. Quality is frequently held up as a supreme good entirely superior to quantity; but nowhere has sufficient folly collected to abandon quantity altogether. In American colleges the quantity unit for academic work is the "Hour" or "Point," and is generally defined as representing one hour (or fifty minutes) of recitation and two hours of outside study per week. An Hour or Point of college work

officially is equal to every other Hour or Point completed with an equivalent quality grade. That is, four Points in a modern language are counted as equal to four Points in chemistry or physics, although the proportion of A's given in languages may be ten times as great as the proportion of equally high marks given in the laboratory sciences. The fact of the matter is that we do not know the value of either of these courses of study, but it is commonly acknowledged that great differences in the quantity and difficulty of work required in nominally equal courses do exist. What is not so well known is that these differences are very large and very frequent, and that they exist within departments whenever there are two or more instructors in the same department.

Most colleges require a certain minimum number of points for the baccalaureate degree. In some colleges practically no common elements are prescribed for a given degree, and in many no distinction is made between the student who completes the required number of points with a grade of A and one who barely gets an average grade of C. It is a common and open secret on every campus that some courses nominally equal to each other are really in the ratio of one half or one third as to real difficulty and amount of work required.

This is the first cause of unreliability in the measurement of college success. Administrators everywhere admit the existence of these inequalities in the Hours or Points of college credit, but do nothing about them for the simple reason that nothing constructive has come out of the criticisms that have been offered. It is hoped that the application and development of objective tests may lead to the acquisition of data sufficiently reliable and meaningful to form a basis for constructive suggestions in the future.

THE BASIS OF COLLEGE MARKS

A second and very potent cause of unreliability in college measurements is the absence of any agreement as to what should be the basis of marking.

Questioning of instructors and administrators in the field brings out the fact that grading as actually done in present practice is, for each individual instructor, based in part or in whole upon one or more of the following factors:

(1) Effort put forth by the individual student.
(2) General intelligence of the student.
(3) General character and personality of the student.
(4) General fitness of the student to live in civilized society.
(5) Amount of improvement of the student—
    (*a*) in general
    (*b*) in specific course or courses, or
    (*c*) in both (*a*) and (*b*).
(6) Actual achievement of the student—
    (*a*) in specific course
    (*b*) in total school situation, or
    (*c*) in total life situation.

To the reader some, perhaps all but one, of these bases for school marks may seem indefensible if not absurd in an enlightened society; another equally competent reader, perhaps, may find this choice the most absurd of all. For example, one college instructor found himself quite unable to believe that there could be teachers "intelligent enough to be on a college staff" who would base their marks upon anything but "the actual achievement of the student in the course and subject matter that is being taught to the student by the instructor who is giving the grade." But to one noted professor and prolific writer on various aspects of education this manner of grading is bad; he complains that "it is even wicked because it so often seems to degrade

or punish those of inferior abilities and to exalt those whose brilliancy is not of their own making." Another professor in the same institution whose contributions to modern educational progress are everywhere recognized, bases his marks almost wholly on the effort put forth by the student and on the growth of the student as measured by the difference between his knowledge of the content of the course at the beginning and at the end of the course; in his own words: "Some students who received a C at the end of the course knew more at the beginning than some other students who received B knew at the end of the course"; in other words, Student X enters the course knowing much less than Y, and ends the course still knowing much less about the course than Y; but X gets a grade of B while Y gets a grade of only C or D, simply because X has progressed a greater distance beyond his initial position than Y progressed beyond Y's initial position.

These instances are not cited because they are unusual; space forbids mention of even a small fraction of the available examples of curious and almost incredible disagreements on this apparently simple matter. The point is that there are very striking disagreements, where at first glance all seemed "satisfactory." A well-appearing, earnest, and amiable, but mentally unendowed student comes to school; he works very hard, he tries so faithfully to overcome his deficiency, and is so piteous in failure that teachers A, B, and C, obedient to one of the theories named above, give him high marks; teachers X, Y, and Z reluctantly but conscientiously report that he has utterly failed to grasp and profit by their intellectual offerings. Yet all school marks of the same outward value are interpreted as signifying approximately equal amounts of the same thing.

There are, of course, arguments for and against each of the attitudes illustrated above. It would seem, however,

highly expedient, in view of the growing social importance of exact information concerning future citizens, to reach a working agreement. Many would say that it is the right and duty of the school administration to know as exactly and as unequivocally as possible the status of each student in each course; it ought to be the right and duty of parents to know just what and how well their children are doing in each of their school subjects; it is the right and duty of society to know what changes are wrought in its children by the various costly units in an increasingly costly educational system. If it is desirable to make record of the effort put forth by the student in each subject matter, this record should be so separate and distinct in its definition as not to be confused by any chance with the actual achievement index in each subject matter; if it seems expedient for each teacher to pass upon the character, personality, fitness for life in civilized society, or general intelligence of the student, these indices should be distinctly defined and kept separate from all other marks. If we are to have grades at all, they ought to be such that they can be used intelligently; the only way they can be made to serve intelligent use is to reach a measurable agreement as to what they are meant to indicate. Grading in college and high school should become more a matter of describing facts exactly and dispassionately and less a vehicle for the personal opinions and moralistic anxieties and predilections of instructors. A grade in Geometry should represent achievement in Geometry and not effort, docility, moral worth, general intelligence, etc.

THE MEANS OF DERIVING GRADES

Over and above the matter of the definition of the intent of grades there is still a very crucial question whose reso-

lution is also prerequisite to a valid system of marking. Agreeing for the moment that effort is the basis of grading, how are we to know, in the very large classes which our mass education makes necessary, the quantity and quality of effort put forth by students? If amount of growth or actual achievement in a given course is the criterion, how are we to judge the status of students in these respects?

Whatever the basis, teachers in general use two methods of ranking their students with respect to that basis:

(1) general personal reaction to the behavior of the student:

(2) written examinations.

PERSONAL ESTIMATE OF THE INSTRUCTOR

This method is far more generally used in substance than is ordinarily supposed, in spite of the fact that its weaknesses have been so clearly demonstrated. It is an individual and subjective mélange of many relevant and irrelevant factors, such as readiness of speech, voice, manner, attitude toward things in general and toward special interests of the instructor, "personality," reputation, and rumors. However absurd it may be, there are few who have not heard stories of the influence which race, color, sex, and skill in the use of cosmetics have on grades derived by personal estimate. The method of frankly direct personal judgment has long been known to be so erratic that many teachers abjure it almost completely, and boast of using only the "fair and impersonal" method of the written examination in which the facts are "there in black and white." This written examination is the backbone of our present system of measuring the products of teaching. A little reflection will show that its "impersonal fairness" is subject to very grave suspicion.

WRITTEN EXAMINATIONS

The written examination in present practice usually consists of from five to ten questions on as many important points of the course in which the examination is given. The questions are chosen by the individual instructor; they are worded by him; their relative value is determined by him, not once, at the time of making out the questions, but as many times as there are papers scored; and the answers to the questions are evaluated by him subjectively. It is difficult to understand how such a scheme could be very much safer than the discredited method of personal judgment. All the vitiations of the latter enter into the former where the instructor is even mildly acquainted with the student; the personal factor is far more powerful and insidious than we are accustomed to think. When the written paper is the sole measure, as when the student is unknown to the scorer, various relevant and irrelevant factors assert themselves. By tradition, the written examination consisting of from five to ten short essays on designated questions is supposed to show the student's "power of organization." What this "power" is, is left to the "good sense" of the individual instructor, as is also the appraisal of any given sample of it. In practice, it would appear that an influential feature of the written examination is sheer mass of reproduction, coupled with any manner of evidence of the student's astuteness in guessing the intent of the questions. The influence of other features, such as physical form of the paper, handwriting, punctuation, and general grammatical form, verbal felicity, rhetorical skill, etc., are generally underestimated. Obviously, even where the basis of marking is exactly defined, reliable measurements are hardly to be expected from such instruments.

It would appear that in last analysis, grades are whatever the individual instructor makes them. In some cases this extreme of individualism is mitigated to a certain extent by various rules (often spuriously, as when authority *ab extra* requires instructors to produce, by means not specified, distributions approximately normal); but in general, grades can be reduced to the unhampered *personal judgment* of the individual instructor at the moment of recording the grade. Some instructors deny, others not only admit but praise this individualism as necessary and just. No doubt it has been the best possible under the conditions of the past; but the acceptance of the traditional methods as adequate for the present stage of civilization rests upon wholly untenable assumptions. Two of these assumptions are so curious that they should at least be named. The first is that English prose is a perfect mirror of a student's achievement and ability; the second is that a person who is an instructor is thereby automatically qualified to give accurate, reliable, and meaningful marks according to modern conceptions of such terms.

THE PURPOSE OF GRADING STUDENTS IN COLLEGE

Of course, it is perfectly proper to ask, "Why do we have grades or measures of college success?" This is something more than a rhetorical question. Some who are convinced of the inaccuracies of the present system and who have witnessed frequent instances of the injustices issuing from such inaccuracies, are in favor of throwing overboard all differentiated grades beyond "Pass" and "Failure" for undergraduate departments of the college, as has already been done in many graduate schools of the universities. For undergraduate departments this might

quite fairly be likened to the problem of the escaping fish who only succeeded in getting out of the frying pan into the fire. Such a division would be a patent misrepresentation of the facts, not even accomplishing the real purpose behind the proposal, namely, to conceal by a very gross classification the inaccuracies thus far unavoidable in the present system. It would seem reasonable to assume that the abilities or achievements of students in the various subject matters are distributed somewhat as the normal surface of frequency; we know beyond question that they are not distributed in two compartments, pass and not pass, good and bad. It would be as reasonable to divide students into the tall and short, the sick and well. There are greater differences among those who pass than among some who are on opposite sides of the passing line.

The purposes of grades in secondary schools and colleges may be divided into two fairly distinct classes:

(*a*) Pedagogical, and

(*b*) Administrative.

(*a*) PEDAGOGICAL

Grades may contribute to a student's education either directly by calling forth his effort in answering the questions and in performing the tasks most visibly used in deriving his grades, or indirectly by forcing him to study effectively in view of prospective examinations, or by enlisting his coöperation through the incentive and competitive possibilities of just and fair marks of his ability or achievement.

(*b*) ADMINISTRATIVE

On the administrative side we may enumerate various plausible reasons for grades in high schools and colleges:

(1) To inform parents or guardians of the educational status of their children or wards.

(2) To give information as to fitness of students for higher or differentiated schooling.

(3) To enable an institution to give a relative standing to students and so determine those to whom credit or degrees or honors should be granted.

(4) To furnish records sufficiently meaningful and accurate to be used in researches looking to vocational guidance and vocational education, to an appraisal or comparison of the efficiency of different school systems, of different instructors, of different methods of instruction, and of different subject matters, and to all questions in the resolution of which exact information on accomplishment is indispensable.

Whether we accept one or all of these as valid purposes of grades, it is clear that each demands a certain indispensable modicum of accuracy now absent.

EXPERIMENTAL EVIDENCE OF THE INACCURACY OF SCHOOL MARKS

A great deal of evidence has been collected which demonstrates the inaccuracies of school marks in general. Only a few bits will be presented here.

Starch and Elliot [1] ('12 and '13) had a final examination paper in first-year high-school English graded by 142 English teachers in as many high schools. If the final examination system for English is valid, teachers using it as an instrument to measure the same identical product

1 *Educational Psychology,* Daniel Starch, PhD., Professor of Educational Psychology, University of Wisconsin. Macmillan, New York, 1920. The last chapter gives an excellent summary of most of the investigations to which reference is here made, together with others not mentioned for lack of space.

ought to agree within reasonable limits, so that the maximum variation would not be more than one- or two-fifths of the total range. We might well endure variations of only one or two letter-grades; but this paper was marked all the way from sixty-four to ninety-eight per cent. Another examination paper of the same kind was rated by the same 142 teachers of English; the scores ranged from fifty to ninety-eight.

A final examination paper in American History was graded by seventy history teachers; one teacher assigned it a grade of forty-three and another ninety; a dozen teachers rated it as eighty or above, and another dozen scored it as below fifty-five.

It might be objected that these are not fair samples of the system because of the peculiar difficulties of measuring English, history, and similar subjects with necessarily indefinite content. These objectors would agree that mathematics is one of the most definite and objective subjects in our curriculum, and that a trial of mathematicians or teachers of mathematics would be fairer. A final examination paper in geometry was scored by 114 mathematics teachers; as in the case of the English and history papers, it was marked on the basis of one hundred as perfect. Twelve teachers marked it as fifty-three or below, and fourteen marked it as eighty-three or above. Comment on variations so great as these is obviously unnecessary.

That these variations in the indices afforded by a time-honored and respected educational device are not due to special characteristics of the high school nor to the fact of geographical distribution of the grading teachers is indicated in an investigation made by Starch ('13), in which ten final papers in Freshman English in the University of Wisconsin were graded by ten instructors in Freshman

English. On one of the papers the maximum variation in the grades assigned by these ten instructors was forty-eight points on a base of one hundred. The average score assigned to the ten papers by one instructor was 65.5; the average score assigned by another of the ten instructors to the same papers was 85.1. Papers in other subject matters were scored by teachers in the respective departments of the same school with similar variations.

A measuring device which in the hands of one competent person gives the length of an object as ninety-two feet, and in the hands of another equally competent person indicates that the length of the same object is only twenty-eight or thirty or forty feet, cannot be relied upon. The well-known compilations of Meyer ('08) at the University of Missouri, of Dearborn ('10) at the University of Wisconsin, of Foster ('11) at Harvard College, and of Finkelstein ('13) at Cornell, all tend to show that the present marking system is thoroughly unreliable even in the hands of undoubtedly competent and careful instructors. In one of these institutions one teacher for a series of five years assigned a grade of A to fifty-five per cent of the students in his classes, and a grade of F to only two per cent; another teacher in the same institution in the same period of time assigned a grade of A to one per cent and F to twenty-eight per cent of his students. At another, one teacher assigned A to thirty-five per cent and F to one per cent, and another assigned A to one per cent and F to thirty-two per cent of his students.

Finkelstein ('13) found that of about 250 Cornell students in the same course throughout the year, but who changed instructors at mid-term, 12.5 per cent were exempted from final examination by one set of instructors, while 37.7 per cent were exempted by the other.

COLLEGE ENTRANCE BOARD EXAMINATIONS IN ALGEBRA AND GEOMETRY

More recent evidence of the inadequacies of the marks derived by the traditional examinations was brought out in a study of the algebra and geometry examinations of the College Entrance Examination Board for June, 1921. About four hundred algebra and an equal number of geometry papers selected at random were scored each twice independently by two different readers of the Board. The correlations between the first and second scorings were very high for both algebra and geometry, about .98 and .96. In spite of this almost perfect objectivity in the scoring, however, the reliabilities of the examinations themselves were very low. The correlation between random halves of the algebra examination was found to be only .61, and that between random halves of the geometry examination only .41. By the use of Brown's formula, the reliability of the whole algebra examination is estimated as not greater than .76, and that of the geometry examination as not greater than .58.

The meaning of these reliability coefficients may be made clearer by a consideration of an hypothetical case closely resembling the actual situation faced by the Board in giving college entrance examinations.

Let us suppose that ten thousand candidates are tested with Form A of a given geometry examination whose reliability is about .60, and that thirty per cent of the ten thousand fail. Now let us suppose that the same ten thousand are tested with another equivalent geometry examination, say Form B, whose reliability is also .60, and which fails thirty per cent of the candidates.

If the reliability of the two forms of the examination were 1.00, the same three thousand candidates would be

failed by both forms; but with a reliability of only .60, the agreement on failures would be as follows in gross numbers:

Failed Form A Passed Form B 1,279	Passed on both Forms 5,721
Failed on both Forms 1,721	Passed Form A Failed Form B 1,279

In other words, accepting the results of one such examination as valid, which was done by the Board in 1921, another equivalent examination would pass 1,279 of the 3,000 failed by the first, and would fail 1,279 of the 7,000 passed on into college by the first.

If we assume that fifty per cent were failed by each of the two forms, the displacements in gross numbers would be:

1,015	3,985
3,985	1,015

The fate of two thousand in ten thousand candidates would be reversed by an equivalent form of the same examination when the reliability is no higher than that of the C.E.E.B. Mathematics C examination for June, 1921.

The causes of the low reliabilities of these examinations seem to be chiefly:

(*a*) The small number and monotonous character of the questions or elements in each; and

(*b*) The inconvenient form of the questions and the consequent time-wasting recording processes.

(*a*) The algebra examination consists of ten questions, only two or three of which give a significant number of partial credits between the limits of zero and perfection, and the geometry examination has only six questions, all of which are similarly defective.

Ten questions do not and cannot constitute an adequate sampling of the algebra field which the examination is supposed to cover. Six questions cannot possibly constitute an adequate sampling of the whole of the materials and mental processes belonging to plane geometry, especially when three of those six are of the same type (the reproduction of Q.E.D. demonstrations from the first book of Euclid).

Six or ten questions do not give an adequate sampling of the candidate's performances. It is a fact upon which there is no disagreement among competent thinkers that pure chance may and frequently does account for the success or failure of a candidate upon a given question. Now if there were one hundred questions, the maximum possible influence of chance on any one question would be only $\frac{1}{100}$ of the total possible score; if there were fifty questions the maximum influence would be $\frac{1}{50}$ of the total possible score; where there are only ten questions, the possible influence of chance on any given question rises to $\frac{1}{10}$ the total possible range, and in the case of the algebra examination here studied, it is $\frac{1}{8}$ the actual range, and more than $\frac{1}{2}$ the Standard Deviation.

When the number of elements is reduced to six, as in the geometry examination here studied, the operation of chance on one single question may affect the total score by $\frac{1}{6}$ the total possible range, which in this case is more than $\frac{2}{3}$ the Standard Deviation.

In general, the smaller the number of questions or elements, the more are all chance vitiations exaggerated.

This is especially true where there is no exact knowledge of the relative difficulties of the elements, and where no effort is made to weight them according to their difficulty or significance.

Such inequalities and inexact weightings are sure to occur in any examination of whatever type; but their influence is greatest when the number of questions is very small, and conversely their influence diminishes rapidly as the number of separately scored elements increases.

If the questions were capable of division to a reasonable extent, all these chance influences would be reduced; but the distributions of the credits assigned to each of the algebra and geometry questions show that more often than not, the candidate either gets full credit or no credit, thus forming what approximates a discrete series. This is one of the strongest objections that has been advanced against the True-False type of question; but the force of the objection is reduced to a negligible quantity when the number of questions is as great as one hundred in any type of test.

(*b*) In a sense, the inconvenient form of the questions is the more fundamental of the two causes here mentioned, because it makes the other causes inevitable: so much time is required in the sheer recording of the responses that it is impossible to give more than from six to ten questions on each examination. The ideal examination is one which uses the whole of the examination period for pure cerebral activity in the field of thought in which the examinee is being measured; at the opposite pole in the scale of efficiency are those examinations which take a large part of the time, energy, and attention of the examinee for purely recording processes and other less relevant activities. The attempt to measure ability or achievement in every branch of science or of the liberal arts through the sole medium of English prose is a notable case in point.

It is a curious and inexplicable fact that even the most intelligent and otherwise cautious teachers attribute to English prose produced under ordinary examination conditions, powers which Roman Senators used to attribute to astrologers, magicians, and oracles. From the confidence with which teachers ordinarily believe that an essay type examination reveals to them the mental powers, special achievement, literary appreciation, organizing ability, etc., of their students, one would easily conclude that such English prose constituted a medium through which the examinee's very soul became as tangible and objective as an Euclidian proof. But this confidence, like so many current non-scientific beliefs, has no better basis than tradition.

For a candidate to write out in good English prose a certain fact, judgment or relation that he has learned in a course of study is one way of showing that he has learned it; but it is not always either the only way or the best way of showing it. Similarly, the writing out of long demonstrations (which in many instances have undoubtedly been merely committed to memory) and the solving of originals in geometry involving tedious arithmetic operations are not the only and probably not the best way of showing achievement in geometry.

Large and unrecognized differences were found between the difficulties of the examinations which were thought to be equal and were used as though they were known to be equal. According to the Reports of the Secretary of the College Entrance Examination Board the percentages of failures in algebra for the years '16 to '21 inclusive have been as follows:

Year	'16	'17	'18	'19	'20	'21
Percentage of failures	61.8	36.7	25.3	61.3	26.1	28.5

These variations are almost certainly due to differences in the difficulties of the examinations used. If we should take five thousand candidates in each of two successive years

from the same geographical areas,
of about the same chronological age,
of the same racial stock,
of the same social and economic classes,
with about the same schooling,
taught in the great majority of cases by the same teachers,
recommended by the same masters and superintendents,
actuated by the same ambition to go to college, and
with equal foreknowledge of the nature and character of the road leading to admission to college,

it would not be within reason to suppose that the true average in any measurable trait or ability or achievement of so large and so similarly selected samplings could vary in absolute quality or quantity more than .5 per cent from one year to another, barring some stupendous and unheard-of cataclysm which by a miracle would profoundly affect the whole population of very large geographical areas.

The passing mark of sixty per cent is one of the interesting curiosities of educational history. Its absurdity is easily recognized when applied to non-educational situations, and its persistence in the most respectable collegiate communities is an enigma to all impartial observers. If a person should act on the hypothesis that sixty per cent of the capacity of one HOUSE is equal to sixty per cent of the capacity of everything else that passes under the same name, HOUSE, we should unanimously agree that such a person was extremely naïve, foolish, and in-

competent; but if a professor makes the same sort of unwarranted assumption about two groups of questions called by the same name, EXAMINATION, his conclusions are accepted without question. Even when attention is called to the fallacy, many conservative teachers argue somewhat as the school architect who insisted on making all the chairs and tables in the sixth-grade class-rooms of equal height on the theory that all sixth-graders should be of the same height.

The distributions of the marks for both algebra and geometry examinations show that the latter are not adapted to the whole range of talent or achievement tested; they are sharply truncated at the upper end, showing quite clearly that the examinations are too easy to differentiate between the students somewhat above the average.

The very grave weaknesses of the customary types of examinations are perfectly apparent when we consider the implications of the data here presented. In the first place, they are based upon experimental results of mathematics examinations, which are the least imperfect of all the traditional types of content examinations. In the second place, these mathematics examinations were constructed, administered, and scored by the College Entrance Examination Board in a perfectly bona fide and official manner, with all its resources of expertness and experience in full play. It is safe to say that the Board has capitalized the uttermost merits inherent in the traditional types of examinations; but the results, as we have just seen from reliable data, are unsatisfactory.

The indication is clear that we can no longer depend upon guessing and individual opinion in the construction of content examinations. We must abandon, or pro-

foundly modify and supplement, the old types of examinations. We must substitute experimental data for guessing, and scientifically tested technique for the traditions of the Middle Ages. How this may be done is described and illustrated in the succeeding chapters.

THE RELIABILITY OF GRADES IN COLUMBIA COLLEGE

That marks in Columbia College suffer from all the vitiations of scholarship marks in general is indicated by all the objective studies that have been made, certain details of which have been presented in preceding chapters. The reliability of indices based upon large samplings of college academic work is shown in Tables 39 to 43 inclusive. These Tables show to what extent college success, in the sense of academic achievement as measured in current grading, predicts itself from one semester to another and from one year to another.

TABLE 39

Semester	Mean	sigma	II	III	IV	(III-IV)
I	14	4.58	.536	.588	.435	
II	14	4.18		.46	.291	
III	14	4.30			.465	
IV	14	4.90				
(I + II)	28	6.70				.632
(III + IV)	28	8.00				

TABLE 39. Intercorrelations of number of points of college credit earned in indicated periods, based upon records of 111 New Plan men admitted to Columbia College in September, 1919, who remained in college two years. No account is taken of quality indices; all points earned with a grade of C or better are counted equally as points earned.

TABLE 40

Semester	Mean	sigma	II	III	IV	(III-IV)
I	16	3.70	.455	.351	.251	
II	16	3.34		.495	.520	
III	16	3.60				
IV	16	4.36				
(I + II)	32	5.85				.536
(III + IV)	32	7.00				

TABLE 40. The facts of Table 39 shown for 122 Old Plan men admitted in 1919 who remained in Columbia College two years.

TABLE 41

Semester	Mean	sigma	II	III	IV	(I-II)	(III-IV)
I	C	2.1	.622	.528	.554		
(I + II)	C	1.44					.706

TABLE 41. Intercorrelations of median college grade for indicated periods of time in Columbia College, based on records of 111 men described in Table 39. No account is taken of quantity in the index here used.

TABLE 42

Semester	Mean	sigma	II	III	IV	(I-II)	(III-IV)
I	105	26.5	.64	.53	.58		
II	115	30.5		.56	.53		
III	105	35.2			.65		
IV	105	37.5					
(I + II)	215	52.6					
(III + IV)	210	65.2				.70	

TABLE 42. Intercorrelations of Scholarship Scores for indicated periods of time, based on records of 111 men described in Table 39. The Scholarship Score takes account of both quantity and quality of work.

TABLE 43

Semester	Mean	sigma	II	III	IV	(I-II)
I	125	30.1	.69	.59	.50	
II	125	33.6		.68	.66	
III	127	33.7			.66	
IV	120	38.0				
(I + II)	250	58.4				
(III + IV)	245	65.6				.724

TABLE 43. Intercorrelations of Scholarship Scores for indicated periods of time, based on records of 122 Old Plan men admitted to Columbia College in September, 1919, who remained in college two years.

These Tables show conclusively two things:

(1) That the combination of quantity and quality indices results in an index that has greater reliability than either the quantity or quality indices taken singly; and

(2) That the reliability of the best combination of present measures of academic achievement is far from perfect—about 0.72.

Taken alone, the quality index is more steady than the quantity index; but on both theoretical and empirical grounds the combination of the two seems better. If the "Hours" or "Points" were more nearly equal, the improvement of the combination would be much greater. But for the present we must look upon 0.70 as the limit of the correlations which may be secured between any criterion for admission and academic achievement as now measured. Until the measures of college success are made more stable, the improvement of admission criteria is definitely halted, since we can already on the basis of a three-hour intelligence examination, predict college success almost as well as it predicts itself. The selection of candidates for ad-

mission may be made more satisfactory by analyzing the elements of college success and by measuring these elements more exactly.

WHO CAN SUCCEED IN COLLEGE?

No one can succeed in college work who is unable to enter into The Great Tradition with a certain minimum degree of ease and appreciation. It might be argued that the athletic and social life of the college is valuable for any one, whether he can respond to the academic environment or not. This may be true, but we cannot be too prodigal with our slender college resources. We can use our resources wisely only by expending them on boys and girls who are best fitted to profit by the distinctive activity of the college. We can lay down the criteria for selecting these individuals best by inquiring first what are the elements upon which academic success depends.

(1) The first requisite for academic success in college, over and above every other prerequisite, is a certain minimum of native intelligence. The lower limit of intellectual powers competent to survive present standards in American and English colleges is not below the twentieth percentile of human intelligence, counting from the top downwards. In Columbia College this limit is not below the fifteenth percentile, and only a few succeed below the tenth. It is possible to graduate from a first-rate high school, even *magna cum lauda*, and still not be a fit subject for the type of work which the college offers. The mastery of secondary school contents is not an insurance policy on the baccalaureate honor.

(2) A certain minimum of preparation, particularly in the case of less intelligent students, is prerequisite to efficient and economical work in college. The extent and

variety of specific preparation is not as great as tradition would have it. The tendency clearly has been to decrease the number of subject matters prerequisite to college entrance, and also to emphasize "power" more than detailed, specific contents. But this tendency can hardly go further if college standards remain what they are now. As long as solid geometry and trigonometry remain the first mathematics courses of the college, we can not, without waste, admit students who are unacquainted with algebra and plane geometry. At least one foreign language must be required and, of course, the vernacular. It has been found that students who cannot pass the high school leaving examinations in the vernacular not only suffer in college English, but also in all courses in which there is any reading and writing to be done, which includes the vast majority of descriptive courses. This triad, mathematics, modern language, and the vernacular, is the minimum for college entrance; it is also the maximum. Beyond this, in our present variable organization of secondary school content, probably the best and only reliable basis for differentiation is the general information factor tested in the intelligence examination.

(3) Character and moral habits are important conditions of college success. Most prominent of the moral habits required is that one of hard work which is all too rare now as in the past. This one is the surest indicator of the others that could be found.

(4) The direct influence of health is so obvious as to require no comment, beyond the fact that the most dangerous dislocations are not those due to sudden temporary illnesses, but rather those due to chronic and probably long undetected and undiagnosed conditions of lowered vitality and discomfort.

(5) Interest is a factor whose neglect in the past is now

being offset to a certain extent by a few scientific researches and by liberal regulations concerning the choice of courses of study by under-graduates as well as by graduates. It is within the experience of every student adviser that many students fail in specific courses not because of a lack of ability to understand the subject matter but because of a lack of interest which often if not always develops into a positive antipathy.

(6) Similar in effect, but different in character, is the factor of adjustment of academic load to capacity, both qualitatively and quantitatively. The happy or unhappy choice of courses of study is more potent in producing success or failure than is commonly realized. Undergraduate deans are frequently heard recounting instances of boys "finding themselves" after one or two years of nearly hopeless failure. Many of these "self-discoveries" are nothing more than the abandonment of programs for which they were not fitted and the election of activities for which they had aptitude. We do not now know enough about these special abilities to afford educational guidance with any assurance, but it is hoped that more exact and objective content examinations may help to discover some of the more obvious misfits earlier than we now do.

(7) Worry and distractions of various sorts frequently ruin a boy's academic record. One of the most painful experiences to which the student adviser is subject is the frequent spectacle of a fine, able, and ambitious youth being driven to very low marks, if not entirely out of college, by difficulties and worry in getting enough to live on while pursuing college studies. Worry over financial matters is hardly more effective in spoiling academic records than the less excusable distractions presented by various campus phenomena from after-dinner hours in

fraternity houses to love-sickness in all its varied manifestations.

Such are the pillars of college success. Most of them are subject to control. If a boy is deficient in preparation, but in other respects has the sort of stuff that college graduates are made of, his admission is merely delayed. If his health is poor, he can be sent to the hospital or a rest center for a season. His financial embarrassment may be relieved in various ways. His character and habits may be modified and improved. His interests may be developed to a certain extent. But if he is stupid, regardless of other qualifications, it is not only useless but harmful and wasteful to admit him to an environment in which he is foredoomed to a cruel, discouraging, and demoralizing failure.

All these seven factors which we have enumerated are important; one or another of them is, as has been indicated, frequently decisive. But native intellect is the *sine qua non* which sets the limits within which the other factors have significance for college work. The preëminent and crucial importance of original mental endowment is plainly indicated by the correlations set forth in the preceding chapters. The nearest rival to the intelligence test in predicting college success is a content examination which approaches most nearly a power test. The correlations of intelligence scores with college success, as has been observed, are far from perfect, not only because of the influence of the other six factors just described, but because of the imperfect indices which we are obliged to use as measures of college success. These measures are not only incomplete but unreliable; they presume to measure only the most characteristic product of the college, that is, academic achievement, and they do that very imperfectly.

It should be clear by this time that there is no real conflict between the intelligence examination and the content examination. There has been a growing consciousness on the part of the colleges of the supreme importance of "power" as opposed to "content." It has become increasingly apparent that mastery of secondary school contents is neither a sure index of ability to pursue college work nor a necessary prerequisite to efficient college work. The specific content that is really required to pursue a normal college program advantageously is not as large or as rigid as was once thought. The main value of the content examinations in the past has been that they have measured intelligence, and have probably measured intelligence better than they have measured content. The weakness of the content examination is its unreliability in measuring both content and intelligence. Psychologically, there is little difference between a good intelligence and a *good* content examination.

The assumptions underlying the policy of those college administrators who insist on formal secondary preparation checked up by the traditional content examinations seem untenable when stated overtly:

(1) That a large amount of formal content is a necessary prerequisite to good college work according to present standards.

(2) That the only acceptable way of acquiring this requisite content is through the formal attendance upon the sessions and the performance of the prescribed work of the secondary school.

(3) That the only way to measure this requisite content is through the traditional content examinations, which remain rather rigid in spite of the so-called "comprehensive" plan.

We have evidence not only of ordinary good sense but

of empirical results to show that a wide variety of formal content is not necessarily prerequisite to acceptable college work. We also know positively that the necessary general information and specific preparation requisite for college work can be gained by means other than secondary schools. We know, finally, that the traditional content examination is a reliable measure neither of content nor of "power" or intelligence.

CONCLUSION

We have shown thus far that the intelligence examination predicts college success very nearly as well as college success, as expressed in terms of the best available index, predicts itself from one semester to another or from one year to another. Since a test cannot predict a criterion better than the latter forecasts itself, all hope of improvement in the intelligence test's prediction depends upon improvement in the reliability and significance of the criterion. We have not found all the causes of the unreliability of our measures of college success; nor have we any very hopeful suggestions for the elimination of some of the causes that we have definitely located, such as the obvious inequalities (both qualitative and quantitative) between various college courses of the same number of credit points and hours of recitation. But a very large part of the unreliability, that due to the indefinite and variable bases of grading and to wrong principles and poor technique of constructing and administering examinations and tests in courses of study, is remediable. The remainder of this book represents an effort to show what the remedy is. The method of exposition will be largely by illustrations in various subject matters and courses common to colleges and secondary schools of superior

rank; but before presenting such concrete suggestions, and for the sake of having in mind a fairly complete and general basis for reviewing and appraising any instrument or method of mental measurement, it would seem desirable to consider first some of the more general and fundamental principles which govern the measurement of mental and social facts. It is with this object that the subject matter of Chapter Seven is introduced at this point.

CHAPTER SEVEN

Some Principles of Educational Measurement

Thorndike has stated clearly the significant features of the conditions which confront us in the measurement of intelligence and achievement. "Measurements which involve human capacities and acts are subject to special difficulties due chiefly to:

(1) The absence or imperfection of units in which to measure,

(2) The lack of constancy in the facts to be measured, and

(3) The extreme complexity of the measurements to be made."

From a consideration of the nature of these special difficulties, he has further stated the essentials of a valid scale for the measurement of mental and social facts. These essentials are treated by Professor Thorndike under the heads, objectivity, definiteness, comparability, and reference to a defined zero point in terms of a defined unit. This chapter is little more than an effort to expand that treatment for the purpose of an elementary exposition. Practically all the material of this chapter is taken from Professor Thorndike's well-known treatise,[1] or directly inferred from some of its propositions.

1. objectivity

"What science means by a perfectly objective scale is a scale in respect to whose meaning all competent thinkers

1 Thorndike, E. L., *Mental and Social Measurements*, Second Edition. Teachers College, Columbia University, New York City, 1912. See Chapter 2.

agree." It may perhaps be easier to understand what is meant by an objective scale if we first consider some scales which are obviously not objective. Beauty is a word which is on the lips of every one; we hear such statements as these uttered every day with the conviction of a martyr's conscience: "This is beautiful; that is not beautiful; this is more beautiful than that." Under the ordinary circumstances of life it is quite unnecessary to go beyond these rather loose terms and ascertain exactly what is meant by them, or what the real facts are which they are supposed to represent. But let us suppose that a prize is offered for the most beautiful work of art produced in a given country in a given unit of time; then it is quite necessary, in justice to all the competitors, judges and donors, to challenge every statement such as "This is more beautiful than that." Except by chance, no two judges would probably agree in arranging the best ten pictures in the world in the order of their beauty. There is no common measuring stick which can be applied to the quality in question; the ratings given by each judge are subjective.

If the problem were to determine the most valuable of two piles of pure gold coins of given denomination, we could either count the coins or weigh the two piles and reach a conclusion which would not be disputed by any competent thinker—the scales of number and weight being fairly objective. But in the case of music it is another story. Granting that the following are the ten best symphonies now known: Beethoven's 9th, 5th, 6th, and 8th; Mozart's Jupiter; Tschaikovski's 5th and 6th; Brahms' 1st and 3rd; and Rimsky's Scheherezade:—would any two competent critics, except by chance, agree in ranking these ten in the order of merit? The same difficulty arises when such words as democracy, justice,

mercy, righteousness, selfishness, etc., are critically examined. Perhaps laymen are not expected to know what justice is; but let it be remembered that not even the judges in our state and national Supreme Courts of Justice agree always as to just what justice is in a given case. The disagreements of art critics are rather more notorious than famous.

The traditional essay examination that is given in most of our courses in high school and college is almost, if not quite, as subjective as the symphony and justice scales just discussed. If a hundred competent teachers of English were to grade one final examination paper in a given course in English, to what extent would their marks agree or disagree? As we have seen, this experiment was actually tried by Starch and Elliot ('12 and '13), with the result that the same identical product was rated all the way from 50 to 100 on a scale of 100. Repetitions of the experiment gave the same or worse results. A final examination paper in American History was graded by seventy teachers of American History; one marked it as low as 43, another as high as 90; a dozen teachers rated it as 80 or above, and another dozen rated it as below 55. Mathematics is a subject matter whose most valuable discipline is its exactness. Geometry is so exact that nothing in human art can do more than approximate its fineness. To what extent do marks derived from the traditional geometry examinations by various competent judges agree? An examination paper in plane geometry was graded by 114 mathematics teachers; one teacher marked it as low as 28, another as high as 92; a dozen marked it as 53 or below, and fourteen marked it as 83 or above.

These variations do not indicate incompetence or lack of honesty and conscientiousness on the part of the

teachers: no one could do more than they have done with scales so lacking in objectivity. To expect grades on essay examinations and the traditional mathematics examinations to be reliable and exact is like expecting guessed areas of passing clouds to be reliable and accurate: it is not certain whether there would be more agreement in guessing the area of passing clouds than in grading a series of traditional examinations. The attack here made is upon the system of grading, not upon those who have been forced, for lack of something better, to use that system.

No system of grading is either absolutely objective or absolutely subjective; in general, no scale of measurement, not even of length or weight, is perfectly objective. In this respect mental scales differ from scales of length and weight only in degree. What we mean by an objective mental scale is merely a scale which is less subjective, about whose meaning there is less disagreement among competent thinkers, than the ordinary scales of mental and social measurement which have been in use. The True-False test is a good example of an objective mental scale. No competent person would disagree in rating a True-False paper, provided they used the key which accompanies the test.

Although objectivity is a necessity in any test, it has no particular virtue apart from the other essentials of a valid mental scale. It is perfectly possible to have a very objective scale without having one which measures the facts to be measured. Thus we might have a True-False test on French pronunciation which would tell us nothing of the relative abilities of the examinees to speak the sounds.

The same remark applies to each of the principles of measurement presented here when taken singly and apart

from all the others. Reliability, for example, is a characteristic without which a test is practically useless. It is usually defined as "self-correlation," and means the agreement of a test with itself. If a test measuring the same object two or more times disagrees with its own findings each time, it cannot be relied upon; either it measures something different each time, or it measures varying fragments of the same thing; either horn of this dilemma is an intolerable situation. Yet it is possible to have a test with perfect reliability, that is, one which agrees with itself perfectly in measuring the same set of facts two or more times, and still not have a test which measures the thing that is to be measured. An extreme example of high objectivity and reliability in a test is the scale for measuring the height of individuals. The scale for height is one of the most objective and reliable scales in the whole field of human measurements, yet there is so little relationship between tallness and mental brightness that no one would think of using the height measure as an index of intelligence. It is only as all the conditions described by these principles occur *together* that they are effective in providing adequate measurements.

2. REFERENCE TO A DEFINED ZERO POINT IN TERMS OF A DEFINED UNIT

Of all the essentials of valid scales, this one is most often absent and its requirements most grossly and mischievously violated. "When any scale of amount is used, one's natural tendency is to interpret it as we have learned to interpret the common scales for number, time, length, weight, area and the like,—to regard every unit in the scale as equal to any other unit and to regard the zero of the scale as measuring *just not any* of the quality in

question. The first necessity in the scientific treatment of any measure is to be aware of the meanings of the units which it represents and of the zero points from which it is reckoned. . . . One must know what fact would, by the scale as used, be measured as zero or just barely above zero. The zero point may be *absolute*, meaning *just not any* of the thing; or *arbitrary*, meaning a point *called* zero, though actually designating some amount of the thing."

In making comparisons the precise nature of the zero point becomes of crucial importance. "Silly as it may appear, we often with mental measurements do such arithmetic as the following: 'John weighs ten pounds more than sixty pounds; James weighs two pounds more than sixty pounds. John is five times as heavy as James.'" This error, stated in this way, seems so obvious that it might be thought superfluous by some to mention it; but so insidious is the real difficulty behind it that even the most careful investigators are not infrequently misled into interpreting *arbitrary* zero points as if they were *absolute*. Just recently the writer was privileged to hear an oral presentation of a Ph.D. dissertation by its author before a group of scientists. One of the major conclusions of that study was just such a mistake as saying that John was five times as heavy as James in the example above. An educational test was given at the beginning of the year to a large number of children and the average score was a certain percentage on that test, let us say, x; the same test was given at the end of the year, and the average score was nearly 3x; from this apparently innocent circumstance the investigator confidently announced the conclusion that the increase in ability in the function measured was nearly 200 per cent! That is, he assumed that a score of zero on the scale indicated absolutely none

of the ability in question; whereas it really designated an amount of ability in the function probably much farther above absolute zero than his second average score was above the first average score.

The history of our present Fahrenheit temperature scale, and its relations to the Centigrade and Réaumur temperature scales, gives us a clear object lesson on the importance of the location of the zero point and of the nature of the units. It was formerly thought that the absolute zero point in temperature, i.e., the point at which all molecular activity ceased, was 32° F. below the freezing point of pure water. Now it is thought that the true and absolute zero point in temperature is 273° C. below 0° C., or 273° C. below the freezing point. Whenever any temperature is announced simply as 59, 15 or 12 degrees we are at a loss to know just what is meant, and we ask almost impatiently, "Fahrenheit, Centigrade, or Réaumur?" That half-impatient question really demands an answer to two questions:

1. What is the zero point from which you reckon the index of temperature?
2. What is the nature and size of the units in which you express that index?

It is important to notice that these questions imply no preference as to what zero point is used or what unit is used. All that is required is adequate definition of whatever point of reference is used and of whatever unit of measurement is used. It would undoubtedly be desirable to know what the absolute zero point is in every scale of human experience, and to reckon all measurements from just that point in every scale; but we do not know exactly what the zero point is in many scales, and fortunately it is not necessary. For all practical purposes it is only necessary that the point of reference be defined

so that any competent person may easily identify it beyond question of doubt, and that the units of measurement be similarly defined.

For example, if it had been specified in the last paragraph that "59" meant "59° F.," we should have known at once that the zero point was 32° F. below freezing, and that the unit of measurement was $\frac{1}{212}$ of the range between that zero point and the boiling point. Knowing these two facts, our experience would teach us that it was weather for a light coat and brisk walking. Or, if it had been specified that "15" meant "15° C.," we should have known at once that the zero point was the freezing point, and that the unit of measurement was $\frac{1}{100}$ of the range from freezing to boiling. Again, and just as certainly as with the Fahrenheit scale, we should have known that it was a case of brisk walking. Similarly, the same certainty and perfect understanding would have been obtained if the Réaumur point of reference and unit of measurement had been used. A real feeling of the importance of exact definition of zero point and of unit of measurement could be acquired by any one who, desiring a bath at body temperature, should hastily act on the information of a careless servant who had tested the water with a Centigrade instead of a Fahrenheit scale and had reported that the water was exactly 98,—really just short of boiling!

It is a question whether it has been or is fortunate or unfortunate for education that carelessness about the definition of zero points and units of measurement in educational scales has not been as fatal as the carelessness of the servant would have been in the case just suggested. How often do we hear instructors of all ranks and in all classes of schools referring to a grade of 30, 50, 70, or 98 as though it were as meaningful as 30° Réaumur or

98° F.? And how much faith is held in the stability of these vague zero points and units of measurement! Just recently the writer heard an instructor say in substance to the Dean of his college, in answer to the question, "Why did student X fail in your course?" "X did excellent work all the term, handed in all written work, recited excellently whenever called upon, passed all monthly examinations, but made only thirty per cent on the final examination." Most probably the final examination was a Réaumur scale which the instructor was using as though it were a Fahrenheit scale, and therefore the thirty per cent referred to would really amount to about 100 per cent when translated into the kind of scale which he had been using all through the session.

The zero point most commonly used in measurements of relationship between scores in different tests and examinations is the Mean or Arithmetic Average of each distribution of such scores. The unit used for locating positions above or below this zero point is a measure of the variability of the distributions, usually the Standard Deviation. The essential quality of a good point of origin and of a good unit of measure is stability. If a zero point wanders like a planet, observations reckoned from it would be as disastrous statistically as observations on a planet whose motion is unchartered would be to the mariner guiding his bark on the high seas by the supposed constant position of that planet. If a unit of measurement, let us say the *inch*, were subject to capricious and unpredictable contractions and elongations, like a measuring-worm or a column of mercury or a rod of steel exposed to various temperatures, objects of equal length would frequently appear unequal as measured by that unit; the same identical object measured twice would appear to have two lengths.

The Mean and the Standard Deviation are not perfectly stable, but they are the most stable point of origin and unit of measurement now available. In measuring the same group with various tests, comparisons may be made directly by expressing all scores as deviations above or below the average in terms of Standard Deviation. But the same test applied to different groups gives both different points of origin and different Standard Deviations. Universal comparisons can therefore be made only when measurements are expressed in terms of the Standard Deviation (and reckoned from the Mean), of some defined and standard distribution, easily accessible to every scientist and investigator. The selection of this standard distribution is a difficult problem whose solution depends upon considerations too complex for presentation in this chapter. In the meantime, the use of the zero point and unit of measurement here proposed, that is, the Mean and the Standard Deviation, even when based upon locally derived distributions, would be such an improvement over the guess-work and unrestricted subjectivism which characterizes our current college measurements that their adoption is to be recommended from every viewpoint.

3. THE FUNCTION TO BE MEASURED MUST BE DEFINED AS ADEQUATELY AND AS EXACTLY AS POSSIBLE

"An ideal scale, such as that for weight, is a series of perfectly defined amounts of a perfectly defined thing, the differences between any two of them being also perfectly defined, so that a series varying by steps of equal difference can readily be selected." This requirement means not only that the function to be measured must be definite and easily identified and recognized by any

competent thinker, but also that the units in which the measurements are expressed must be definite and easily recognized by any competent thinker.

A moment's reflection on the extreme complexity of any mental function will show the importance of this principle. There seems to be an irresistible tendency on the part of the vast majority of human beings to assume unconsciously that they know all about that of which they know the name merely. Ask any normal person, "What is reading?" and if he is not actually insulted by the apparent reflection on his intelligence, he will tolerantly but confidently define the term as it would be defined in a much-abridged dictionary; ask a philosopher the same question, and he will meditate long and seriously. It is a vast darkness and ignorance that is screened for most men by the word "reading" and its synonyms. "Reading" is one word, and for the average man it signifies one definition. But it is only necessary to call attention to the fact to convince any normal person that reading is many things masquerading under the appearance of indivisible unity, and that the recognition of its complexity is fundamental even to the most elementary definition that may be given of it.

Certainly no measurement worthy of the name can be made of reading which does not somehow recognize and take account of its complexity. Efficiency in reading may be likened to the efficiency of an automobile. The inefficiency of an automobile might be due to one or more of a vast variety of unitary or complex causes, such as bad headlights on a dark night, punctured tires, poor lubrication, exhausted gasoline supply, poor gasoline, poor road, etc. Now a test of total objective and effective efficiency under any given conditions is valuable if that is all that is desired; but that would be a measure

only of a result and would not in itself give any hint as to the cause or causes and possible remedies. In mental measurements the chief interest is usually if not always centered in the problem of improvement, and hence the aim is not merely measurement of results but to measure in such a way as to show the causes in analytic form. In measuring reading ability in a given individual we do not merely wish to know how well he reads under the given conditions, but also whether and to what extent his reading is influenced by bad printing, bad lighting, near-sightedness, far-sightedness, astigmatism, the vocabulary of the material used, the language of the test, the interest of the individual in the test, and the mental capacity of the individual.

The necessity for complete definition of a function to be measured does not imply that the measurement of complex functions should be avoided as unsound or despaired of as impossible of achievement. The principle of definition means simply that the complexity is to be recognized where it cannot be analyzed and that our ignorance of its analysis must be borne in mind, not only in the construction and administration of a measuring scale, but especially in interpreting the results of such measurements. Far from avoiding the measurement of complex functions, exact definition will always increase our power to cope with complexity, and show us more clearly the necessity for measuring total complex functions. We might measure the efficiency of a large number of the individual elements in a complex and still know little or nothing of the effective efficiency of that complex.

The importance of exact definition is apparent in any scientific discussion of any educational measurement. For many people, for some even in responsible educational positions, geometry is geometry, reading is read-

ing, history is history, and algebra is algebra. The tendency to confuse identity of name with identity of fact is not confined to the ignorant masses. Measurements of various things called "geometry" which correlate with one another little better than pure chance events, have been accepted and acted on by school administrators with a confidence and authority little becoming an age of science. Aside from the failure to appreciate the variety of elements which are really included in the term "geometry," elements which would have appeared clearly in any careful definition, a more fundamental error has been committed by these authorities who pass judgment with such celerity: they have failed to recognize the distinction between *ability* and *achievement*. Fortunately, the two functions are highly correlated in most cases; but it is well-known that some high scores on the traditional geometry examinations are due to real geometrical ability, and that others are due to much industry either on the part of the student or of the teacher over a very long period of time. Further treatment of this very important distinction is reserved for a special section farther on. It is mentioned here simply as a glaring instance of the baneful effects of inadequate definition in matters of measurement.

The definition of the function to be measured in college and high school courses here advocated is "knowledge of, and ability to think in, the material of the course."

4. CONSISTENCY

"The series of facts used as a scale must be varying amounts of the same sort of thing or quality." Just what this means in a specific scale depends upon the

definition of the function that is to be measured by that scale. It does not mean that a scale may not consist of a series of complex things or qualities, but that it must throughout be composed of the same defined things or qualities, be they complex or unitary. If the thing or quality whose amount it is desired to determine is a complex consisting of equal amounts of x, y, z, and w, then the series of facts used as a scale for the measurement of that thing or quality must consist of equal portions of x, y, z, and w at *all* points on that scale.

If we desire to measure achievement in a given college course and we define that achievement as "knowledge of, and ability to think in, the material of the course," then in so far as our examination in that course measures such things as speed and quality of handwriting, proseform, personality, voice, etc., however desirable excellence in these matters may be, it falls short of being consistent with its avowed purpose. And if our examination is of the traditional subjective kind, which is mainly a measure of the personal predilections of the scorer for a variety of elements, consistency may be considered as almost entirely absent. It is as if we attempted to measure the specific gravity of some light substance like cream by ascertaining how far the ordinary instruments sink in a sample of salt-saturated water which has a very thin layer of cream on its surface.

A notable example of obvious impurity of measurement is afforded by some arithmetic tests which are, unfortunately, still given in many of our supposedly modern school systems. Some of the problems in these tests are but little more than very severe reading tests. If reading ability of a very special kind is a part of arithmetic, the tests are probably sound; but in spite of the desirability of reading it would seem more advantageous for all pur-

poses of measurement to separate the two functions. As given in the past the tests have been incapable of revealing conveniently what particular difficulties have kept particular individuals from making satisfactory progress. A valid scale must measure the thing it is supposed to measure throughout the whole range of difficulty or quality that it covers; we must not imagine that it will measure the function we are interested in merely because we give the scale a particular name.

5. THE VARIABILITY OF MENTAL FUNCTIONS IN THE SAME INDIVIDUAL FROM DAY TO DAY AND HOUR TO HOUR MUST BE RECOGNIZED IN THE CONSTRUCTION, ADMINISTRATION AND INTERPRETATION OF ALL TESTS AND RESULTS

We must not only avoid such loose and undefined terms as "ability to add," especially in dealing with young students of arithmetic; we must not only always specify ability to add one-, two-, or three-place numbers, etc.; but we must specify whether the index of such ability is derived from one, two, three, or ten or *n* trials. It would be interesting to attempt to enumerate the causes of variability or inconstancy in mental facts, if that were possible; but our chief concern here is whether there *is* such variability. Even in comparatively simple and apparently unitary functions, such as adding or subtracting one-place figures, the achievement of individuals per minute varies quite noticeably. A typical distribution of the performances of an individual in adding two-place numbers follows:

Score:	7	8	9	10	11	12	13
Frequency:	1	3	6	5	3	1	1

The insecurity of accepting results of one trial as a basis for estimating the true average ability of an individual even in a very simple function is apparent from this figure. The more complex the function tested the more insecure is any estimate based on a single trial. Certainly we should not attempt to estimate an individual's ability to read, to solve algebraic problems, to demonstrate geometric propositions, to spell, etc., on the basis of his reading of one sentence, his solution of one problem in factoring, his demonstration of the parallelism of chords perpendicular to the same diameter, or his spelling of one word.

The principle of increasing accuracy accompanying the increase in number of items of a test is illustrated in a case of coin flipping. Thus, if ten coins are flipped, the chances are even that the number of heads will be within one of five; that is, within about twenty per cent of half the number of coins. If 100 coins are flipped the chances are even that the number of heads will be within four (theoretically $3\frac{4}{10}$) of fifty; that is within seven per cent of half the number of coins. If 1,000 coins are flipped, the chances are even that the number of heads will be between 489 and 511; that is, within about two per cent of half the number of coins. If 10,000,000 coins are flipped, the chances are even that the number of heads will be within .00067 of half the number of coins—almost exactly half.

Mathematically speaking, we would say that the relative variability of the number of heads decreases in proportion to the square root of the number of coins flipped. Now in all probability the reliability of a test varies with the increase in the number of questions in much the same way. This means that, making four times as many questions would give us about double the reliability; making

nine times as many questions would give us about three times as great reliability, etc. The conclusion is that scores on examinations in every subject matter must depend on as large a number of trials as it is possible to give the examinee. As was pointed out in the last chapter, ability or achievement in plane geometry cannot be measured with six questions. A good spelling test should consist of at least 100 carefully chosen words. A test in ancient history should give the examinee at least 100 chances to win or lose. Only by taking a large sampling of an individual's performances can we arrive at a reliable estimate of his normal or average ability of the kind tested.

6. COMPARABILITY

A scale is valid only if it can be applied with reasonable ease and accuracy to the objects or facts to be measured. Calipers are better for head-length than an ordinary footrule because the caliper-scale is more directly and accurately comparable with the thing to be measured. We may have means of making exact chemical analyses of all substances in our laboratories, and yet be unable to analyze the contents of the moon, the sun, or other distant objects by means of the reagents in our laboratories. But other means of analysis may be as applicable to the contents of the sun as to those of the earth on which we live. The spectrum analysis is in constant use, and has already been the means of learning *something* of the chemical components of our sun, and also of some of the more distant suns. Such indirect methods in the physical sciences already have counterparts in the methods of mental and social measurements. This topic is discussed further in Section 9 of this Chapter.

7. THE DISTINCTION BETWEEN "POWER" AND "ACHIEVEMENT" MUST BE KEPT IN MIND IN THE CONSTRUCTION, ADMINISTRATION, AND INTERPRETATION OF ALL TESTS AND TEST RESULTS

A specific illustration which came under the observation of the writer suggests itself. Two children about seven years of age were measured for ability to read; the score of A was indefinitely below the average for her age, that of B slightly above average. When this became known, the teacher put A in the slowest beginning group, and B in an advanced group in the second grade. The Terman intelligence test revealed the fact that A had an I.Q. of 130 while B had an I.Q. of 97. Investigation showed that A had grown up in a negligent rural family and had never had an hour's instruction in reading, and that B had grown up in New York and had been in school for over three years, including kindergarten. The writer thereupon urged the teacher to put A in the most rapid beginning group; after much argument, the teacher reluctantly complied. To make a long story short, this reluctant teacher promoted A to the second grade at the end of four months, and at the end of the year A was more than one P.E. above B in ability to read as defined by the Thorndike Alpha II. The difference between achievement and power is clearly indicated.

8. ANY INFERENCES AS TO THE GENERAL INTELLIGENCE OR INBORN ABILITY OF TWO INDIVIDUALS MUST BE BASED UPON THEIR REACTIONS IN MATERIAL TO WHICH THEY HAVE BEEN EQUALLY EXPOSED AND IN WHICH THEY HAVE HAD EQUAL PRACTICE, EXCEPT INSOFAR AS EXPOSURE AND PRACTICE ARE INFLUENCED BY NATIVE CAPACITY

This introduces an objection frequently made against intelligence tests by the uninitiated. During the War army officers condemned the Army Alpha Intelligence Examination because it "was not a measure of capacity but of achievement, not a measure of intelligence but of educational opportunity and social environment." Not even the makers and propagators of the Army tests deny the partial truth of the assertions of such critics. The answer of the psychologists is not that they did not fail to take account of the principle stated above, but that they failed by so small a margin as to disarm critics. That they failed by a very small margin is attested not only by a detailed examination of the test but by experimental results. The most striking evidence was found in the fact that many soldiers who had enjoyed the best educational and social advantages made lower scores than many soldiers who had been reared in rural districts and had attended only the poorly taught and poorly equipped schools which characterize American rural districts. In some cases soldiers who were bona fide college graduates made lower scores than some country boys who had never passed the sixth grade!

Our interest here is not a defense of the Army test; these facts are cited because of their bearing on the exposition of the principle under question. The question is not specifically, Did the Army test succeed; but, How may the conditions laid down by the principle above be most clearly fulfilled? In the first place, a preponderant emphasis should be placed upon testing mental processes which are largely independent of informational content, such as solving simple reasoning problems which increase in difficulty by fairly small and regular increments, or appreciating relations and making judgments in verbal material with which all English-speaking Americans are

perforce familiar. In those tests in which the answer depends markedly upon special experience, it is not necessary that every individual tested shall have been exposed equally to every element, but that each individual shall have been exposed to at least an approximately equal percentage of the special experiences involved in the test. This is well exemplified in the so-called information test which is used in the rôle of a power or intelligence test. Every such test in the Army examinations included equal proportions of items of information from about eight fairly distinct fields of normal human experience in America, such as agriculture, science, literature, history, sports and athletics, biography, etc.

Under normal circumstances it is usually necessary or convenient to consider that the same number of years in school constitutes "equal exposure" to the content and experiences connoted by the curriculum; but for the sake of stimulating constructive criticism, it is just as well to remember that this is a mere assumption, which in many cases is certainly not true. Of course, large differences in exposure may occur, as between the curricula and teachers of different schools, or in the same school; but even in the same school, in the same class, in the same subject matter, large differences in actual effective exposure may and do exist. Equal exposure does not depend on external conditions only: the interest, incentives, and distractions that condition each individual's experiences and modify and direct his efforts are perhaps the most potent of all the factors which influence the quantity and quality of exposure to the varied material of school experience. The home and social environment are all-powerful in strengthening or weakening the bonds which are set up or torn down by formal education. Differences in past experience at the time of exposure may make large

differences even when all other factors are equalized. Strictly speaking, only when all these influences have been accounted for can differences in achievement be strictly attributed to differences in general intellect or special abilities.

9. WHERE INDIRECT MEASUREMENTS HAVE DEMONSTRABLE ADVANTAGES IN SIGNIFICANCE OR RELIABILITY OVER DIRECT MEASUREMENTS, INDIRECT METHODS SHOULD BE USED

This principle is immediately suggested by the discussion in the last paragraph. The information test when used as a contributory test of intelligence, or as an achievement test in a specific high school or college course, is often criticized unjustly by those little conversant with its theory and use. The criticism is usually based upon a fallacious distinction between "mere facts or bits of information," and "reasoning, organizing ability, synthetic power, etc." "Reasoning capacity," "organizing ability," etc., are subject to the weakness of very imperfect definition as used by the critics of the "information test"; but some light is thrown on their meaning by the fact that they are supposed to be measured quite satisfactorily by the traditional essay examination.

The essay examination in the hands of the average teacher does measure a very important element which apparently cannot be measured directly by any other means thus far developed. But it measures that element very incompletely and very unreliably. It seems impossible, even under the most careful conditions, to obtain a reliability greater than 0.65 or 0.70 with the essay examination. This is due fundamentally to the inade-

quacy of the sampling which it affords of the examinee's performances and of the material on which he is being tested, and to the variability and uncertainties of subjective scoring.

Indirect measurement is not suggested as a substitute for, but as a supplement to, direct measurement. The information test is highly reliable not only because of its objective scoring but because of the adequacy of the samplings which it affords of the performances of the examinee and of the material on which he is tested. Moreover, it is flexible enough that it can be made to test not only "information" but also judgment, and the appreciation of relations, causes, and consequences. Even if it tested only "information," with its greater reliability it would be an invaluable supplement to the direct measurement.

There is not as much opposition between "information" and "reasoning" as some teachers would have us believe. Facts do not exist in the mind in isolation. We remember by thinking, and we think by remembering facts. Those who declaim against tests of "mere facts acquired and remembered," should realize that we cannot think without facts. We could not neglect the measurement of facts without neglecting the measurement of the very fabric of thinking. When we consider that facts are not only a legitimate and undoubted aspect of thinking, and also that they can be acquired, retained, and reproduced only by thinking, only by organizing material in a logical and systematic manner, there can be no doubt of the value of a pure information test. In fact, the measurement of reasoning and organizing ability in a field of knowledge would be very defective and incomplete without a measurement of the breadth of information in that field. Breadth of information cannot be tested with the essay

form of examination in a three hour period, for the simple reason that on physical grounds only so many pages of prose may be produced in three hours; and even if this defect were absent, range of information could not be objectively and reliably measured with the essay form of examination.

Fortunately, we do not have to depend on theoretical considerations. Every experimental study thus far made and reported has shown a very high relationship between measurements of information in a field and intelligence or ability to think in the material of that field. A stupid individual *may* be able to retain and reproduce a large mass of more or less detached facts in a given field without any broad synthetic understanding of that field; but these cases are so rare and usually imply so many other desirable qualities, such as industry, persistence, patience, etc., that they may be disregarded in this discussion. The data presented in the next chapter show that the "information" test agrees with itself in ranking a large group of students. On both theoretical and experimental grounds, then, indirect measurements are a valuable supplement to direct measurements.

10. A VALID TEST MUST CONTAIN A LARGE NUMBER OF SMALL ELEMENTS, OR, IF THE UNITS OF THE TEST ARE LARGE, THEY MUST BE DIVISIBLE IN THE ASSIGNMENT OF CREDITS

This principle at once permits an adequate sampling of performances of the individual tested and of the material on which he is tested. To fail on one question in ten is not equivalent to failing on an equal percentage of 100 questions; to pass on nine out of ten is not equivalent to passing on ninety of a hundred. Ability

or achievement in any given subject matter, such as geometry, algebra, spelling, history, etc., cannot be measured reliably with one question, nor with ten or even twenty. Traditional examinations have consisted of from five to ten questions or problems which are to be answered each by a brief impromptu essay, or by fairly large bits of complex and highly organized mathematical reasoning. This means that only a few spots in the subject matter can be touched upon; it means that many pages of prose or manuscript must be produced under particularly urgent and exciting conditions; it means that all answers must be expressed through a medium whose use is in itself no easy task for the average examinee; it means that the scoring of the answers will be highly subjective and little susceptible to standardization; it means that many irrelevant factors will influence the score, such as handwriting, spelling, grammatical form, prose style, etc., and, where the student is known to the scorer, such factors as personality, voice, color of hair, of eyes, of skin, age, race and sex, docility, adeptness in flattery, etc. All, or nearly all, of these weaknesses in the traditional methods of examination are overcome by the apparently simple principle of "many small elements."

When we use a large number of small elements, chance influences are reduced to a minimum. As we increase the number and decrease the size of the credit elements in a test, subjective scoring approaches objective scoring, and progressively achieves all the important secondary effects of objectivity. A familiar illustration is afforded by the traditional examination in algebra or geometry. In spite of the apparent objectivity of geometrical demonstrations, they are little susceptible to objective evaluation. It is difficult to judge what went on in the mind of the examinee from what he wrote into his demonstra-

tion. One scorer might conclude, from a very small irregularity in the arrangement of supporting axioms or previously proved propositions, corollaries, etc., that the examinee failed utterly to understand the proposition in question; another scorer may judge very leniently a fundamental error. The maximum possible error which a scorer may make in such cases is obviously the total value of the question in the examination. If there are ten questions in the examination, and all are accounted of equal value, the maximum possible error which may arise from the subjective scoring of one question is ten per cent of the total possible range of scores on the examination; and where, as frequently happens, sixty is the passing mark, the influence of one subjectively scored question may be twenty-five per cent of the maximum range of those who pass the test successfully. If there had been a hundred questions, other things being equal, these maximum possible errors due to the subjective scoring of one question would have been, respectively, 1 per cent and 2.5 per cent instead of 10 per cent and twenty-five per cent.

11. IN THE MEASUREMENT OF ANY MENTAL FUNCTION AS MANY TYPES OF QUESTIONS SHOULD BE EMPLOYED AS ADMINISTRATIVE CONDITIONS ALLOW

This principle partly overlaps, partly grows out of, all the preceding ones. Its soundness and importance are indicated by a variety of considerations. It affords a chief means of meeting the requirements of Principle 3, above, concerning the definition and complexity of functions. Wherever the traditional essay type of examination is used exclusively, this principle is violated. English prose is one means of expressing or con-

veying certain kinds of ideas and thoughts; but it is not the only means, and in many cases not the best. It is an injustice alike to the examinee and to the examiner to force the former to communicate all his ideas and thoughts and judgments through the medium of English prose alone, and to compel the latter to base his evaluation of the examinee's ability or achievement on the prose that can be produced under the conditions of current examining. In college courses of the descriptive type the essay examination is clearly inadequate; it tests an important element, but only one important element, and that very imperfectly, through a medium incapable of being made objective in its results. It does not test breadth of information, ability to make fact-bound judgments throughout the whole range of the content of the course, or the appreciation of many types of relations which form some of the most important aspects of descriptive courses, cultural and technical.

A notable case of the violation of this principle is apparent in the all-too-recent custom of measuring spelling ability orally, in the old spelling-bee fashion, or in writing, the words being dictated without context. These methods are not quite worthless, as some enthusiasts have supposed; they are merely incomplete, measuring only one aspect of spelling ability.

The time-honored examinations in geometry are perhaps the best illustration of the violation of the principle under discussion. The typical geometry paper consists of six questions, three of which are Q.E.D. propositions, and three mixed problems, involving applications of geometrical principles to practical situations. A moment's reflection will show any one conclusively that such a test violates all the principles heretofore discussed, but most prominently the one now under consideration. One of

the things to which geometry is supposed to contribute is facility of clear and logical speech and writing; another is the ability to appreciate and state clearly and succinctly exact definitions and relations; another is the ability to think geometrically without figures and without formal proofs,—to geometrize on one's feet, as it were; another is the ability to construct and draw figures, with and without problems and proofs. But in such examinations we find not a line of original prose required, not a single definition demanded, no thinking without cut and dried proofs (which usually have been memorized verbatim), no reference to the vast amount of geometrical material not sampled by the six questions.

Another reason, quite independent of the argument for complete tapping of complex functions, imposes this principle upon us. Any form of test is impure in the sense that it includes and measures elements which are foreign to its purpose. This largely follows from the fact that nearly if not quite all mental measurements are indirect to an appreciable extent. For example, no one can doubt the large and insidious influence of handwriting, style, etc., in the essay examination; and even if we could so train readers that they would be always absolutely indifferent to all such influences in so far as they were irrelevant, we should still measure to a certain extent such things as speed of muscular movements necessary for writing, differences of emotional reaction to the situation of free prose composition under the urgent conditions of competitive examinations, etc. It is, of course, clear that these are only marginal forces. It is this marginal character which enables us to minimize and cancel them out by using multiform questions, such as the True-False, Completion, Recognition, Association, Trade-form, etc., all of which are recommended on both

a priori and experimental evidences for all descriptive courses, and many others.

12. IN ANY KIND OF EXAMINATION, THE QUESTIONS SHOULD INVOLVE AS LITTLE AS POSSIBLE IRRELEVANT CONSIDERATIONS AND SUPERFLUOUS ACTIVITIES ON THE PART OF THE EXAMINEE

Stated positively, this principle simply means that every question should be constructed and set up in a form that fulfills all the following conditions, or effects the best possible compromise between them that is possible in view of the ultimate purpose and immediate conditions:

(*a*) It should cover fully and unambiguously the goal aimed at;
(*b*) It should involve no unnecessary writing;
(*c*) It should contain no unnecessary distractions;
(*d*) It should contain no "trick" or puzzle elements of an illegitimate character (e.g., double negatives);
(*e*) It should lend itself to objective scoring methods;
(*f*) It should allow rapid scoring;
(*g*) It should be little subject to chance influences of any kind.

The first condition requires no discussion. The second is so simple that it is almost a presumption to call attention to its necessity; but it is so frequently and unnecessarily violated that some illustrations seem appropriate. The essence of this principle is economy of time allotted for measuring purposes. This economy can be effected only by increasing the proportion of our examination periods that is given over to thinking on the part of the examinee, and *pari passu* reducing the proportion that is given over to purely mechanical and irrelevant tasks involved in the recording of responses. If we ask, "When

and where was Shakespeare born, and when did he die?", we would normally get, on the traditional essay examination, at least one sentence: "Shakespeare was born at ——— —— ——— on the — day of——— in the year ——, and he died in the year ——." Suppose that instead of asking the question, we should simply write out the answer we desire, leaving blank spaces in which the student can write the data omitted. We could then in the same time also test several other facts about Shakespeare.

"His remarkable literary career came to an end in the reign of_____when England's international prestige was highest and when such literary figures as_____, _____, and_____in England, _____and _____in France, and_____and_____in Spain, and_____in Germany were in the full vigor of maturity. His most noteworthy contemporaries were _____, who wrote_____,_____who wrote _____, and famous as the most prolific writer of _____that the world has known."

This illustration also serves for the other conditions set forth above. One of the main distractions and annoyances and time-and-energy-consuming superfluities of the essay examination is the production and writing of prose. This example not only reduces the amount of writing to a minimum, but thereby removes a very large and annoying distraction to effective and free thinking; it contains no tricks or puzzles; it is capable of almost absolutely objective scoring; it rescues so much time from irrelevant activities that a broader sampling of subject matter is possible; and it reduces the influence of chance to a minimum. Simplicity and rapidity of recording responses must not be confused with simplicity of the thought processes necessary for arriving at the correct responses.

The recording of the answer to a Completion, True-False, or Recognition question is a very simple matter; but the thought processes leading to that response probably would not in many cases suffer by qualitative comparison with those involved in writing brief essays.

13. A VALID EXAMINATION MUST BE A TRUE TEST FOR THE WHOLE RANGE OF ABILITY OR ACHIEVEMENT WHICH IT IS SUPPOSED TO MEASURE

No absolute criterion is available to show whether an examination fully satisfies this condition, but fairly secure indices are not wanting. Undistributed scores anywhere, especially at either or both ends of a distribution are almost a sure sign that the instrument is not valid for the whole range of talent which it is designed to differentiate. When the frequencies pile up at the upper end it means usually that the test is so easy for many of the examinees that it does not differentiate them. It is as if they were measured for height with a measuring stick which by some strange means would not register differences in the heights of individuals above seventy-one inches. When the frequencies pile up at the lower end of the scale, we have the situation which would be given in height measurements if our measuring stick failed to register differences in the heights of individuals below sixty-six inches, and called all heights at or below sixty-six the same.

To illustrate in terms of mental measurements, let us suppose that we have ten individuals that differ from one another by uniform amounts on some definite scale of ability, so that their respective true scores are from lowest to highest, 10, 20, . . . 90 and 100. Now if we should make an examination of ten questions, the first

so easy that all of the ten individuals could answer it correctly, the second easy enough for the nine best to answer successfully, the third not too difficult for the eight best, and so on up to the tenth, which is made so difficult that only the best one could answer it; then we should have an ideal scale for these particular ten individuals. But let us suppose that the first three questions were left out of this examination; then the first, second, and third individuals would all get the same grade, that is, zero, because none of the questions in the examination was within the range of their powers. Thus we have undistributed scores which ought to be distributed if they are to represent the real facts measured: the first three individuals differ one from another just as much as the other seven men differ one from another; but because the examination is too difficult, they all get the same score. If the three most difficult questions had been omitted, the three best individuals would have received the same score as the fourth best, since there were no questions difficult enough to allow these three men to show the real differences between their abilities. Many illustrations might be cited from reports of examinations in our school systems.

An examination must not only be consistent in the sense of measuring the same thing throughout its range, but it must be sensitive to and capable of registering real differences in every part of the range of the quality it is designed to measure.

14. THE ELEMENTS OF AN EXAMINATION MUST VARY IN DIFFICULTY BY APPROXIMATELY EQUAL AMOUNTS

This is necessary for several reasons, among which are the avoiding of undistributed scores, and the insuring of satisfactory reliability of measurements. In the example

above, let us suppose that the second and third questions in the examination were just as easy as the first. Then the first three individuals would all get the same score of thirty; or let us suppose that the fourth, fifth, sixth, and seventh questions were all as easy as the third, then all these individuals would get the same score as the third. Of the ten individuals, one would get a score of ten, one twenty, five thirty, one eighty, one ninety, and one one hundred. There would be greater differences in real ability among the five who got the same score, than among some of the individuals placed in different groups by the examination. All who could pass the third question, but could not pass the eighth, would get the same score.

15. CHANCE INFLUENCES MUST BE RECOGNIZED AND COUNTERED IN THE CONSTRUCTION, SCORING, AND EVALUATION OF EVERY TYPE OF QUESTION

This caution is stated as a principle mainly to remind us of the larger and more systematic chance factors that are inherent in many types of questions. In the True-False tests, in which half the statements are correct and half incorrect, any ignoramus might answer fifty per cent of the questions correctly by marking them all as true or all as false. Where the number of alternative answers is only two, pure chance would always tend to give a score of fifty per cent correct answers. The defense against this chance influence is to score every True-False test by the "Right-minus-Wrong" method. By this scoring method pure chance would always tend to give a score not significantly above or below zero. But the use of negative scores in compensating for chance influences which are neither systematic nor measurable is scientifically indefensible.

In the Recognition type of question with four alternatives, pure chance would tend to give every examinee a score of twenty-five per cent correct responses. Where five alternatives are presented, the chance factor is reduced to one-fifth instead of one-fourth. This can be further reduced by making some of the alternatives in each question very seductive. The best use of this form of question seems to be to require every student to answer every question, and to take the "Number Right" as the score. This would very nearly equalize the chance factor as between the various individuals taking the test, and thus preserve with high reliability their relative standing.

In many tests, incomplete, vague, or variable directions introduce large chance influences. The directions should always be clear and uniform. Where large comparisons between different groups taking the test are to be made, differences in the conditions of administration, including the physical layout, must be taken into account. Practice influences the score obtained on every form of test, and in making comparisons between individuals or groups their relative familiarity with the types of questions used must be taken into account. In the intelligence test special experience may frequently produce spurious results if the choice of materials and of types of questions has been narrow.

16. IN THE CONSTRUCTION AND ADMINISTRATION OF EXAMINATIONS MEASUREMENT MUST NOT BE CONFUSED WITH PEDAGOGY

This does not mean that good measurement is incompatible with good pedagogy; but no pedagogical end can be a valid excuse for poor measurements when accurate, reliable, and usable measurements are desired. It has

been asserted that the efficient forms of question here suggested have no pedagogical value. This is an independent question whose answer does not affect the status of such forms of questions as regards measurements. Teaching is one thing, measuring another; when we wish to teach, let us not assume that a device which has pedagogical value also has measurement value, or that a device which has little or no pedagogical value has no measurement value. A strong case might be made for the pedagogical values of the new types of questions if that were necessary. Many students and instructors have commented upon the fact that the True-False question, for example, forces the examinee to be very critical in reading any given statement, and tends to develop a habit of exact use of the King's English. But this apology is entirely beside the point when we are interested in exact measurement. Pedagogically, anything that makes students feel that they are under fire or on trial is a desirable stimulus, perhaps, whether it measures anything accurately or inaccurately, especially if the exercise has intrinsic pedagogical value. But certainly from the viewpoint of stimulus, depending in part upon the idea of competition, the more exact measurement has the advantage, regardless of its intrinsic pedagogical value, because no student will have any deep respect for marks which are known to depend upon subjective individual judgment and to be unreliable or unfair. Subjective, and therefore unreliable systems of school marks, have encouraged many students to seek to "make a good impression" on their teachers by various non-academic devices. Intrinsic pedagogical value in an examination is highly desirable, but the value of the examination as a measuring device cannot be made to depend upon its value as a teaching device.

SUMMARY

Our principles have been presented under sixteen heads, but it must be remembered that these divisions are made mainly for convenience of exposition. No one of these principles has any particular merit except when conjoined with all the others. The overlapping and interlacing of them has been apparent throughout. We cannot too often repeat that all such principles or rules of measurement are only devices designed to overcome the absence or imperfection of units in which to measure, the lack of constancy in the facts to be measured, and the extreme complexity of these facts. Illustrations of the application of these various principles in the same examination are reserved for succeeding chapters.

It may not be out of place to make a remark here on the interpretation and use of exact measurements. Accurate measurement must be clearly separated in our minds from administrative policy and social philosophy. The only legitimate purpose of a measuring scale is to classify individuals reliably, and place them in relative positions accurately with respect to a defined trait. The height measurements of recruits for the U. S. army are, from the viewpoint of measurement, absolutely impeccable; the scale of height used fulfills all the conditions of the principles in this chapter. But these measurements do not tell us what recruits are too low or too tall to become soldiers. That is a matter of policy; and just as each of the great military powers decides what height a man must show in order to become a soldier, so must each college decide just what degree of stupidity will be tolerated above the admission or "passing line." It is quite common for instructors to imagine that an examination can and should tell us just who should fail and who

should pass in a given course, and thus relieve us of the responsibility of wielding the academic club. What actually happens with the traditional examinations is that in accordance with local policy or, more often, on the subjective opinion of the instructor, a "passing line" is established arbitrarily and called "60" or "70." With the new types of examinations, the definition and application of educational and administrative policy still rests upon those who use the examinations.

A final caution, which is really a principle of science in general, is that opinions on the relative values of different types of examinations should, wherever possible, be based upon objective measurements of their reliability and significance, rather than upon tradition, *a priori* considerations, vested interests, or general impressions gained from off-hand observations.

CHAPTER EIGHT

Experiment with a New Method of Content Examining [1]

PLAN AND FORM OF THE NEW CONTENT EXAMINATION

Researches in connection with the use in Columbia College of the Thorndike Intelligence Examination as one of the criteria for admission brought out among other things the fact that grades given to students in Columbia College are subject to very grave charges of inaccuracy. In a suggestion to the Dean that the matter be experimentally investigated it was pointed out that Columbia College presented in the course Contemporary Civilization A1-A2 ideal conditions for a thorough-going trial of various kinds of measuring devices. This course, since it is prescribed for all Freshmen, includes a large number of students of varying ability, studying the same material as indicated in a very detailed syllabus, in the same time, by the same method of instruction and recitation, and under the direction of an unusually harmonious staff of instructors. In addition, there was available a reliable objective mental classification of the students in the course, as well as other data indicative of what might be expected from them.

The experiment as proposed was to include three steps:

(1) An agreement as to what was to be measured;
(2) The construction of a scale as nearly according to the principles described in the preceding chapter as possible under the circumstances;

1 The writer makes grateful acknowledgment to Mr. H. E. Buchholz, President of Warwick and York, Inc., Baltimore, for his kindness in permitting the use in this chapter of certain material contained in an article published in the September, 1921, issue of *Educational Administration and Supervision.*

(3) To measure every student in Contemporary Civilization with that scale, independently of and without in any way interfering with the regular examinations and gradings of the various instructors.

The suggestion, accompanied with a sample of the New Examination, illustrated below, was passed on to the instructors in Contemporary Civilization, who, in accepting responsibility for conducting the experiment, made it plain that their action was not to be interpreted as an endorsement of the New Examination. This fact is mentioned as evidence that the experiment was not a biased attempt to find support for a pet scheme or to force evidence into preconceived directions. While the experiment was proposed by several members of the staff, and all were quite willing to undertake the laborious task in the interests of science, the group of instructors as a whole presented quite frankly and obviously a skeptical attitude as to the outcome.

Random Sampling from the New Examination Questions. The questions below have the numbers which they had in the various Parts and sections of the New Examination, and will serve to illustrate the characteristics of the New Examination, particularly the great range of material that can be covered in a short time by questions of these types.

PART I

Directions for Part I. Read these statements and mark each one at the left of its number, with a plus sign if you think it is true, with a zero if you think it is false. Each statement marked correctly gives you a credit of one point; each incorrectly marked statement counts as a penalty against you, and is subtracted from your score; omitted statements count neither for nor against you. Your score will be based upon plus signs and zeros; don't waste time writing anything else. First, go through the list quickly and mark all that you know for certain at once; then go back and study out the harder ones. Do not guess. The chances are against you on guessing. A wrong response counts heavily against your score. Don't endanger your score by gambling on those questions about which you know nothing.

A

9. The civilization of the Phœnicians, Greeks, and Romans centered about short river valleys.
12. Irrigation is carried on more extensively on the eastern than on the western slopes of the mountains of California.

B

2. Education of the right sort increases the innate capacities of students.
16. Among animals under conditions of domestication instinctive behavior tends largely to be displaced by habitual or acquired modes of behavior.
17. In learning how to do something, e.g., playing the piano, the effective responses need not be actually performed in order that the appropriate connections within the nervous system may be made and become habitual.
49. One advantage of customary morality is that customs adapt easily and quickly to changed conditions.
6. Mercantilism was the result of the nationalistic commercial rivalry which had supplanted that of the Italian and German city-states.
18. The enclosing of the "Commons" brought about a great increase in the number of small holdings.
28. Voltaire admired experimental science and "natural law," and was an ardent critic of contemporary conditions.
36. The success of the American Colonies in throwing off the British yoke is the more remarkable since they were opposed by a united and powerful England.
51. By his conquests in Central Europe, Napoleon retarded German unification by at least a hundred years.
68. The Declaration of the Rights of Man declared in favor of private property.

PART II

Directions for Part II. Fill in the blank spaces so as to make sense and convey accurate and significant information. Each unbroken line stands for one word, but you may use more words if necessary.

A

2. The uneven distribution of natural resources makes South America largely dependent for its ..... on the United States; it makes the textile mills of Great Britain dependent for their cotton supply largely upon ..... and .....; it renders the German iron industry al-

most helpless without the mines of .....; it leads the ..... interests of the United States to be concerned about the Government of Mexico.

B

4. The causes for physical fear are largely ..... under the normal conditions of civilized society. But fear commonly displays itself in many subtle forms. In the economic life, for instance, we find that a ..... percentage of laborers are in more or less constant fear of .....; and even where conditions are more stable, they are in fear of descending in their .... In politics fear is responsible for much of our ....., and not only in politics, but in our intellectual life in general we are prone to avoid the ..... as though it were dangerous.

C

6. By the Commercial Revolution the civilization of ..... was extended to the .....; the standard of living was raised by the influx into Europe of ..... from ..... of ..... from ..... and of ..... from .....; and political power was shifted from the ..... aristocracy to the .....
8. By the mercantilist policy the European powers sought to increase their ..... and to decrease their ..... Especially the ..... of raw materials and the ..... of finished products was encouraged. A representative mercantilistic statesman was .....
11. At the Congress of ..... in 1814-15, the foremost ..... of Europe assembled. In the name of ....., France was saved. In Italy ..... supremacy was replaced by ..... supremacy. In the Germanies a ..... was created with ..... at its head. Belgium and ..... were ..... and put under the House of .....
13. The doctrine of checks and balances was advocated by Jefferson as a safeguard against the oppression of a .....; but by John Adams as a safeguard against the tyranny of .....

PART III

Directions for Part III. Underline the one word or phrase in each line that will make the statement most sensible and true. You are not penalized for errors in this test unless your choice is obviously

preposterous, so that it pays to take a chance even where you are not quite sure.

A

5. Most of the inhabitants of the British Isles belong to the race known as the Alpine, British, Teutonic, Mediterranean, Basque, Celtic.

B

1. Rural isolation is made unsatisfactory by the instinct of—Gregariousness, Love, Sympathy, Imitation, Reflection.
7. Miserliness or kleptomania are abnormal exaggerations of the instinct of—Hunger, Love, Acquisitiveness, Fear.
9. Under the régime of a truly reflective morality man's moral choices would be predominantly—Voluntary, Emotionless, Habitual, Customary, Instinctive.

C

4. The Commercial Revolution by the 18th C. had established the commercial supremacy of—Spain, Venice, Portugal, England, Germany, Holland.
7. The followers of John Wesley are known as—Quakers, Baptists, Methodists, Unitarians.
11. The Tudors owed their power to the fact that their policies were in the interest of the—Working class, Nobles, Small freeholders, Bourgeoisie, Church.
21. Jefferson believed that the United States would continue to retain their democratic form of government so long as they—Promoted industrial development and the growth of city population. Remained an agricultural people. Upheld strong centralized government instead of the petty subdivisions of wards, towns, etc. Prevented any tampering with the constitutions designed by the Fathers.
30. The passage of the Reform Bill of 1832 was a victory for the—Agricultural landlords, Farm laborers, Factory workers, Well-to-do middle class.
31. In his political philosophy John Locke advocated—Divine Right, Social Contract, Communism, State Socialism, Economic individualism.
38. The period from 1815 to 1848 is especially significant because of the—Ultimate success of the Metternichian principles. The checking of the power of the Bourgeoisie in favor of the lower classes. The sweeping religious reforms within the Christian churches. The gradual development of the principles of the French Revolution.

THE MAKING AND ADMINISTRATION OF THE NEW EXAMINATION

The original sample included four kinds of tests—True-False, Completion, Recognition, and Association. The staff decided to try out the first three by giving two of the three hours of the final examination period for the purpose. A committee was appointed to draw up the New Examination, and to make up an examination of the Old or Essay type to take the first hour of the three-hour final examination period. It is believed that this One-Hour Essay Examination is superior to the average essay type examination that is given in colleges.

Contemporary Civilization A1
1920-1921
Final Examination
One Hour

Directions: Choose ONE topic from EACH group. These topics cannot be fully discussed on the basis of any one section of the syllabus. To discuss your two topics successfully, you must bring together and organize material from various parts of the semester's work.

GROUP I

1. Discuss conservatism or opposition to change, including in your discussion the bases of conservatism in human nature, the influence of various types of geographic environment making for conservatism studied this semester (as, for instance, in the manorial system, gild system, etc.).
2. Discuss science, scientific method and the laws of science, including in your discussion a definition of scientific method and an analysis of the part science has played in the intellectual, economic, and political changes from medieval to modern life.

GROUP II

1. Define the major groups or classes in England or in France in the early 18th century, and show in some detail the conflicts of these groups and the changes wrought in them and their relative position by the political, economic and social changes of the 18th and first half of the 19th centuries.

2. Discuss the relative position and condition of the Maker of cotton or woolen goods in England before and soon after the Industrial Revolution, bringing to bear upon this discussion your knowledge of social conditions and of human nature.

The general procedure followed by the committee was to assign the questions to the various large divisions of the course covered up to the time of the examination approximately in proportion to the time and emphasis that had been spent upon each such division. Groups of such divisions were then assigned to the several members of the committee; the questions were constructed by each committeeman by reading through the syllabus and through the texts and selecting important facts and relations; those that seemed best suited to the True-False form were tentatively put in that form. When each committeeman had collected a large number of questions on his assignment, they were presented to the full committee, which revised and culled them carefully. This detail is noted here because a combination of judgments in the making of any kind of examination is highly to be recommended: it often happened that a very poor or indifferent question was made over into a very excellent one simply by changing a word here or correcting an ambiguity there which had escaped the originator. In this way the committee collected and presented for the approval of the staff about 250 questions of all kinds. The One-Hour Essay Examination was approved as shown above. Of the New Examination questions, the staff approved 139 True-False statements, 14 Completion paragraphs, and 61 Recognition type questions. These were printed on standard 8x11 paper, and bound into a booklet convenient for handling and for recording the responses easily and unambiguously.

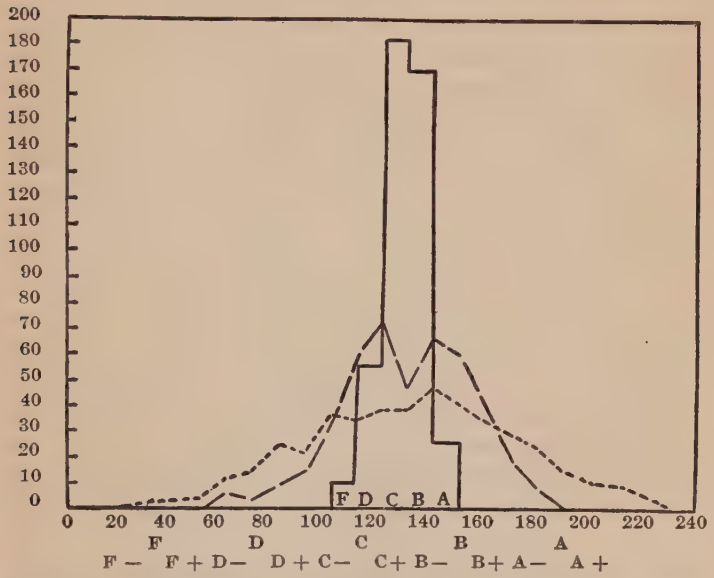
There was some fear lest the number of questions was too great for a two-hour period, but this fear was quickly

dispelled. The test was given to 440 freshmen in Contemporary Civilization at one time in one room; fully a fourth finished the test and left the room before time was called; one finished in fifty-eight minutes. The booklets were scored with the aid of stencils and keys which were prepared by the whole staff sitting as one committee.

One of the first and most significant questions asked by the various instructors was, What will be the passing mark on this test? Most examiners and schools cling to "sixty" as the passing mark, and fondly imagine that a definite and invariable standard has been set up. In practice this scheme simply amounts to a combination of two very mobile variables: the specific examination used, and the individual instructor scoring it. The procedure is based on an assumption which is contrary to fact: we know that some examinations are to others as a rod is to a yard, and that therefore sixty on one may be better than eighty or poorer than forty on another. The only common element in the traditional examination is the useless one of having nearly always 100 units; the relation of these units to each other, in the same or in different examinations, is usually a matter of pure chance or of guessing.

Obviously it would have been absurd to fix beforehand the passing mark on the New Examination, because no one knew how difficult it was. If the traditional spurious sixty per cent correct had been adopted, more than half the students would have failed, since the average score was about 125 out of about 250 possible points of credit. In this experiment the controlling factor in evaluating the New Examination scores was the practical desire of making them comparable to the existing standards in the course. This was done by securing a distribution of grades derived by all means other than the New Exami-

nation, and then fitting the distribution of the New Examination scores to that of the other grades by making the Means of the two coincide, as is roughly indicated in the figure below. The New Examination scores



Graph showing types of distributions of college grades given to the same group of students in the same course. The narrow curve in the center is the typical distribution found in nearly all college courses where five degrees of merit are in vogue. The dash line curve shows the great improvement due to the use of plus and minus signs with the five letter grades. The dotted line curve shows the distribution of the same men effected by the New Examination.

were then expressed in terms of the variability of their distribution, translated into letter-grades, and averaged with the old scores in order to derive the final marks of the students for the semester.

The reader has probably already surmised the basis of grading adopted in this experiment: Knowledge of,

and ability to think in, the material of the course as presented in the syllabus and in class-room meetings.

STATISTICAL RESULTS

The data used in the statistical studies summarized below include the following:

(1) Thorndike Intelligence Scores. These were secured from the Admissions Office, and are based on the Thorndike College Entrance Intelligence Examination which constitutes part of the admission machinery of Columbia College.

(2) Instructor's Estimate of Intelligence. Each instructor gave an intelligence rating to each student in his class on the basis of all available data save the Thorndike scores. These ratings were expressed in terms of letter-grades, A meaning very superior intelligence, B superior, C average, D inferior, and F very inferior.

(3) Instructor's Estimate of Achievement in the Course. This rating was secured in the same way as (2) above, being based upon all data available to each instructor on each student in his class, including class work, weekly and monthly quizzes, mid-semester examination, etc. These ratings were made before the final examination was given.

(4) Score on the One-Hour Final Examination of the traditional essay type.

(4a) Re-score on the same by a different instructor.

(5) Part I of the New Examination, consisting of 139 True-False questions, the score being the number of Right Responses minus the number of Wrong Responses.

(6) Part II of the New Examination, score on fourteen Completion paragraphs.

(7) Part III of the New Examination, consisting of

sixty-one Recognition questions, the score being the number of right responses.

(8) Total New Examination Score (= Sum of Parts I, II, and III).

(9) Combination of (3) and (4) above, by simple average of the two scores. In order to secure a more workable distribution, all the ratings supplied by the instructors were in terms of the conventional letter-grades A, B, C, D, and F, each divided by the use of plus and minus signs, e.g., A +, A, A —, B + B, B —, and so on to F. —.

(10) Combination of (4) and (4a) above, by simple average of the two scores.

TYPES OF DISTRIBUTIONS

A very important fact appears at a glance in the distributions of these ten series of data. All the available data on the students involved in this experiment indicate that their real abilities and achievements are distributed approximately as the probability curve. Space allows only one example. On the One-Hour Final Examination about eighty per cent of the students receive a grade of B or C; only about four per cent get either A or F. This means that practically eighty per cent of the students are unclassified, though there are quite sensible differences among them, as shown by the fact that when the artificial restrictions are removed through the use of plus and minus signs, the classification is much improved, the number of significant steps of merit being increased to eight or ten. The distribution of New Examination scores approaches the normal surface of frequency even more closely, giving about fifteen significant steps in a total range of twenty steps.

TABLE 44

Pearson Coefficients of Correlation

	1	3	4	5	6	7	9	11	12	13
1. Thorndike Intelligence Scores.										
2. Instructors' Estimate of Intelligence	.421		.630							
3. Instructors' Estimate of Achievement	.358		(.707)[1]							
4. First Grading of One-hour Essay Exam.	.386		(.663)							
5. New Examination, Part I	.451	.557	.605	(.800)	(.761)	(.733)		.625		
6. Part II	.497		.611		(.800)			.626		
7. Part III (Rights)	.447		.586			(.573)		.621		
8. (R-W)	.442		.567							
9. New Examination Total (III R)	.513	.620	.654				(.905)	.680	.533	.623
10. New Exam. Total (III R-W)			.644							
11. (3) plus (4)										
12. Second Scoring of One-hour Essay Exam.										
13. Combination of First and Second Gradings of Hour-Essay Exam.										

1 Coefficients placed in parentheses are reliability coefficients based on 117 random cases; in all other cases, except those numbered 12 and 13, N = 325.

The important fact illustrated by Figure 1 is that the instructors themselves, when allowed to do so, classify the men in such a way as to give a distribution sensibly similar to the normal curve. The traditional five-letter scale is so hedged about by regulations and traditions that in reality it permits only a two- or three-fold classification with any freedom. As to the A-grade, it is reserved only for the elect few, for the very superior and his too-few companions; D and F are grades reserved for the other extreme of the curve of success, for the obvious failure as a college student; for the great mass of students between these two wide extremes there remain only two categories, B and C. This is a notable example of the failure of educational administration everywhere to appreciate and use fully the information which instructors are able and usually willing to give, even with the very imperfect instruments of measurement which have been available.

CORRELATION RESULTS

The correlations derived from the data mentioned above offer very strong evidence in favor of the New Examination as a method for measuring school work. We know from various sources that the more accurate school marks become, the more highly they correlate with general intelligence up to a certain limit, say, about 0.50 or 0.60. Any one of the three parts of the New Examination correlates more highly with the Thorndike Intelligence scores than any of the marks derived by the old method of assigning grades; and the total of the New Examination gives a considerably higher correlation index than either the One-Hour Essay Examination or the Estimated Achievement.

But it would be erroneous to charge that the New Examination is therefore simply another intelligence test. It correlates still more highly with the One-Hour Essay Examination and with the Estimated Achievement. These two coefficients are almost as high as that between the One-Hour Essay Examination and the Estimated Achievement, the latter being .70, and the former .65 and .62, respectively. Part I alone gives $r = .605$ with the One-Hour Essay Examination, while Part II alone gives $r = .611$, and Part III, $r = .586$. The near-equality of these coefficients is significant with regard to the reliability of the New Examination, and will be recalled when this important matter is discussed farther on.

In view of the careful manner in which the Hour-Essay Examination and Estimated Achievement grades were derived, they may be accepted as more reliable and accurate than is usual with similarly derived marks as ordinarily dispensed by teachers. It is not a violent assumption to argue that the average of these two marks is more accurate and reliable than either one alone; granting this, we have another check on the New Examination, for if it is a good instrument, its measurements will correlate more highly with the average than with either one alone. Not only does the New Examination Total correlation with the average rise, from .65 to .68, but the correlations of the three parts of the New Examination with the average rise slightly.

An interesting observation may be made at this point as to the relative merits of the three parts of the New Examination. It will be noticed that Part II, the Completion test, gives the best results by whatever criterion it is judged; Part I, the True-False questions, holds its rank as second best by whatever criterion it is judged; Part III, the Recognition test, holds third place by all

the criteria. But of even greater significance is the near-equality of the indices of the three Parts, because this indicates great reliability in the New Examination. The slight differences may be due to the time element, since Part I was allotted seventy, Part II thirty-five, and Part III only fifteen minutes. The efficiency of the New Examination method is evident from all these considerations.

Additional evidence that the New Examination index is limited by the weakness of the criterion is afforded by data on 117 random booklets of the One-Hour Essay Examination which were regraded in lots of about ten each by twelve members of the Contemporary Civilization teaching staff. The booklets were so arranged that those who regraded them could not know the student's name, his instructor, or the first grade on the booklet. Presumably, the average of two judgments on the examination booklets is better than either one of the two judgments alone; and if the New Examination index is limited by the weakness of the first grades on these booklets, the correlation index with the average ought to rise. The correlation of the New Examination with the first grades on the 117 booklets is $r = .599$; with the re-grades, $r = .533$; and with the average of the two, $r = .623$.

INFLUENCE OF PERSONAL ACQUAINTANCE

It may not seem out of place to interject at this point a remark on the influence of personal acquaintance of the marker with the student whose work is being judged, since it has been asserted earlier in this paper that personal factors have a large influence. The first grades on the 117 booklets described in the last paragraph were given in each case by the student's own instructor in the course; the regradings were, as pointed out above, entirely

incognito. The correlation coefficients given in the last paragraph indicate that the essay examination gives better results on the whole when the student is known to the reader; but specific cases are available to show that the influence of this personal factor is so great and so insidious that it leads the most careful readers astray. A booklet graded A minus by the student's instructor was regraded as D; a B was lowered to D minus; a C plus was raised to A minus; a C was lowered to F. The regrading increased the variability from $S' = 2.06$[1] to $S'' = 2.84$, and placed the Mean about .41 $(S' + S'')/2$ below the Mean of the first grading. It may be inferred that whenever the quality of an examination paper disagrees noticeably with the reader's existing personal appraisal of the student, objectivity of scoring, in whatever small degree it may be said to exist in such scoring, simply vanishes altogether: the reader who knows the student and has formed a good opinion of him will unconsciously (or perhaps consciously) overlook and "explain away" the weakness of his examination paper; the reader who does not know this bright-appearing and easy-tongued student, and thereby has no temptation to become a court of equity, grades his paper according to the standard applied to the other papers.

RELIABILITY OF NEW AND OLD METHODS

On the count of reliability all available data favor the New Examination method. By reliability, as has already been stated, is meant the amount of agreement between the measurements of a series of objects given by the same measuring instrument on two or more different occasions. The correlation between the Estimated Achievement

1 S' and S'' are respectively the standard deviations of the distributions of first and second scorings.

grades and the One-Hour Essay Examination grades is $r = .70$; but because of the undoubtedly large influence which the estimate of general achievement had on the grading of the One-Hour paper, we can hardly consider it a valid reliability index. It was to get a straight reliability coefficient on the traditional essay examination that 117 booklets were regraded. The correlation of the first grading with the regrading is $r = .663$. This, although it almost certainly flatters the essay examination method considerably, is a little less than the correlation between the New Examination and the average of the Estimated Achievement and One-Hour Essay grades. The New Examination correlates with the Hour-Essay Examination almost as well as the latter correlates with itself! The self-correlation of Part I alone of the New Examination is $r = .80$, taking even so rough a division as alternate fives of questions; the reliability of fourteen short completion paragraphs of Part II, seven odd-numbered against seven even-numbered, is $r = .80$; and the reliability coefficient for the whole of the New Examination, by the same rough grouping of questions, is $r = .905$. By applying Brown's formula, $r_{22} = \frac{2r_{11}}{1 + r_{11}}$,[1] these three coefficients become 0.89, 0.89, and 0.95, respectively.

GENERAL CONSIDERATIONS

Although the correlation results furnish the most scientific and convincing evidence about the New Examination method, there are certain general considerations the appreciation of which happily does not depend upon familiarity with statistical terms.

At first sight, the New Examination appears to put a

1 In this formula, r_{22} is the reliability coefficient of the whole examination and r_{11} is the reliability coefficient of one half the examination, that is, the correlation between one half and the other half.

preponderant emphasis upon sheer memory of isolated facts; but a study of the questions reveals that, on the whole, they constitute a severe test of judgment and of thought-organization in the material of the course. The questions involve facts, more facts than can possibly be touched upon in the old essay type of examination, but they do not call for a mere memoriter reproduction of disconnected facts. We cannot have thinking without facts. Modern psychology teaches us that facts do not exist in isolation. Except by chance, which is provided against in the great number of questions and in the manner of scoring, a high mark can be secured only if the student has organized the "isolated" facts into a definite system of thought.

The large number of facts which can be touched upon by this method is possibly the strongest point in its favor. One of the chief weaknesses of the traditional essay examination is that it has, and must have, so few questions—usually about ten. Many instructors feel this limitation acutely; they quite reasonably become impatient with the idea of determining a student's gains from four or five months' study in a course by having the student prepare in three hours eight or ten short impromptu essays on a few leading questions. The New Examination is not restricted in its scope to a few "high spots" in the course; it tests the student's knowledge of and ability to think in the whole content of the course.

The large number of questions is made possible by the simplicity and efficiency of the process of recording responses. This is not to be confused with simplicity of the thought-process. The thinking and judgment involved in responding correctly to each of these questions would probably not suffer by qualitative comparison with the processes by which the brief essays are produced.

In the essay examination a comparatively large proportion of time is devoted to the mere recording of what the student has thought out; in the new method a minimum of time is required for recording responses, and a maximum is left free for unhampered thinking.

Nor is there any significant distraction in the new method of recording responses, as in the customary examination. In addition to the distraction and time lost due to the purely physical elements of producing several pages of manuscript, there is quite an appreciable distraction and loss of time in casting the responses in agreeable prose form. Up to a certain point, the essay examination is really a test of English composition ability in a peculiarly urgent and impromptu situation. Whether this is a desirable feature or not is a matter for the philosophers to determine; but it is well to recognize that there is a mixture of functions, which in many cases of special disability undoubtedly produces serious distortion of indices.

Some instructors and many students feel that the very limited number of questions possible in the essay examination renders it subject to frequent and grave chance errors. It is certainly far from absurd to say that many good students might by pure chance know little or nothing about one, two, or even three questions in a given series. In a ten-question examination, where all the questions are considered of equal value (which is very frequently the case), such a chance situation would, at one fell stroke, reduce a student's grade by one, two or three-tenths—a reduction dangerously near the fateful point of sixty and failure. The New Examination, by reason of the large number of questions which may be included in it without crowding the examinee for time, reduces the factor of chance to a minimum; the student has two or

three hundred chances as against ten to pass or fail. To fail on ten questions out of a hundred is not the same as to fail on one out of ten.

In still another way the number of questions has considerable influence on the validity of an examination. While it may be only in primitive communities that examinations are made of eight or ten questions which are *assumed* to be of equal value and of equal difficulty, and are marked each on the basis of one-eighth or one-tenth of the total possible score, still it must be admitted that where differences in the values of questions are considered at all, they are invariably appraised by purely subjective methods. Errors inevitably creep in wherever and whenever subjective methods are used for such purposes. It is not claimed that errors of evaluation of questions are avoided in the New Examination. But it is apparent that in examinations which have only a few questions, such errors of chance and subjectivity enjoy their maximum influence. In an examination which consists of 200 or 300 questions, there may be very large differences without significant vitiation. Where the scale units are very gross, as in the essay examination, all manner of chance or systematic error in evaluation and in scoring is exaggerated to the utmost.

It will be recalled that the first essential for a valid scale for mental measurement was stated by Thorndike to be Objectivity. The facts of a subjective scale are well illustrated by the regrading of the essay papers mentioned above, and by the data given in Chapter Six. Such variations are impossible with a scale such as the New Examination. Only in the completion and association parts does the subjective element appear at all; and here the units are so numerous and small that possible distortion is reduced to a minimum. In all other parts the

scoring is absolutely mechanical, and may be done by intellects of lesser weight than the expensive ones now periodically forced to the unwelcome task of reading hundreds if not thousands of pages of very poor prose. Whereas equally competent readers vary as much as fifty per cent or sixty per cent of the total range of credits in judging the same identical examination paper of the essay type, it is inconceivable that readers could vary more than a maximum of one or two per cent in grading a given example of the New Examination.

PEDAGOGICAL VALUE OF NEW AND OLD METHODS OF EXAMINATION

A thoughtful student of the problem has suggested the question of the comparative pedagogical values of the New and the Essay types of examinations. He says that we "frequently hear the thesis that a quiz is only incidentally a means of marking; that its primary importance is its pedagogical value. It gives the student an opportunity to develop logically a topic or theme, to practice selecting relevant facts, to react in an orderly and systematic way toward problems," etc.

This question, which relates essentially to the proper functions of examinations, is very opportune. The theory of this experiment is that accurate marking is a desirable objective entirely independent of pedagogy, and that accurate measurements of school products are becoming more and more indispensable in a complex society in which the problem of adjustment and adaptation daily becomes more acute. The purpose of the experiment was to demonstrate the value of the New Examination method *as a measuring device.* There is really no controversy possible over this question, so long as we recognize that

there are two distinct ends to be sought, two distinct functions to be served; and so long as we recognize that each may be served better by different agencies or methods. Nothing in this report can be interpreted as an attack upon any pedagogical device whatever; but the fact that a type of examination has pedagogical value does not necessarily imply that it also has measurement value.

The writing of essays is a time-honored pedagogical method which is in a sense the historical descendant of the medieval disputation; its vogue in modern education in the rôle of a pedagogical device is attested by the frequency with which we find the work of courses in high schools and colleges described as "lectures, conferences, (*written*) *reports, theses, assigned papers,*" etc. While the pedagogical value of themes, theses, and dissertations is unquestioned when used solely as a teaching device, there is grave doubt as to how much of this value survives the special conditions which obtain when the essay is used ostensibly as a "test." The old disputation was primarily a test of ability "to think on one's feet"; the essay examination is largely a test of ability to write in a hurry, under compulsion, as ideas come. When thus used it is a serious question whether it is not really negative in its pedagogical influence: it certainly often leads to the careless use of English, and may sometimes encourage hasty and disordered thinking.

ECONOMY OF THE NEW METHOD

While the principal economy of the New Examination method is its greater accuracy and reliability regardless of cost, the question of time and labor required for effec-

tive administration of measuring devices is distinctly favorable to the New method. Though the questions cannot be made out in the same time required for making out a good essay type examination, there is an enormous saving of time and labor in the scoring of results. The mimeographing or printing of the New Examination questions will require only a negligible expense of time and labor which will be more than offset by the economies of objective scoring.

These conditions hold especially for large classes, and for examinations whose main purpose is that of measurement. For very small classes, and for quizzes whose main purpose is pedagogical or disciplinary, the New method is not especially recommended from the viewpoint of economy in time and labor, although in point of accuracy and reliability it would be better even for such small groups and special purposes.

No technical knowledge of psychology or statistics is required for the administration of the New Examination method beyond what may be easily learned by the average instructor in an hour's study. The casting of a distribution of the scores, the calculation of the median and other percentiles, and the assignment of letter-grades to each percentile range, are matters which may be thoroughly mastered by doing them once under the guidance of a statistician, or by consulting any elementary textbook on statistics.

CONCLUSIONS

One is justified in concluding that a combination of the New and Old methods affords for the present the safest starting point. The best procedure is to look

through the syllabus of the course (or whatever serves in its place) and select important points or facts or relations or judgments to which the student should react in a certain way, and then make up for each item so selected just that form of question which will elicit that response from the student in the most efficient way if he knows it. Only two precautions are necessary; first, insure high reliability by including not less than 100 questions in all; second, take the Mean or Median as the point of origin in evaluating scores. It is important to remember that the purpose here is to set forth a method, not a "standard" examination. The New Examination does not of itself tell just how many or which students should be passed or failed; but it does put them in preferential order: it ranks the students in order of merit with respect to the subject matter represented by the examination. The instructor still will be required to decide from his own judgment just where the ax should fall, just as he now does with the traditional method; but with the new method he can make his judgment more intelligently, and with greater accuracy and reliability in the final outcome.

The new method, which may involve True-False, Completion, Recognition, Judgment, Association, or Trade-form questions, is adaptable to nearly all descriptive courses, including all the social and physical sciences. It is especially adaptable for testing knowledge of all foreign languages, ancient and modern. Except for courses in English composition, in which, of course, the essay is the peerless instrument, the new method lends itself easily for examination purposes in all English and literary courses. The best test for mathematics, as for English composition, is a performance test; but the principles of the new method can be applied even here, most especially in the case of mathematics examinations.

INSTRUCTORS' OPINIONS ON NEW METHOD

It may be of interest to some readers to know the attitude toward the New Examination of some of the instructors who sat in judgment on it in this experiment. Some of these instructors have written memoranda for this purpose, of which the following are excerpts. These statements are all the more significant when we realize that their authors in some cases were at the beginning quite skeptical of the new method.

1. My experience with the New method has thus far been too brief to warrant a categorical statement, but all the evidence (and of this I am reasonably convinced) seems to point to the conclusion that it is much the most accurate method of testing students' knowledge of the subject that we have yet discovered. It does not, however, in my judgment, test sufficiently the student's power of accurate and cogent expression, and his ability to organize his material. I should therefore feel that in an examination adequate time should always be given for at least one or two questions of the essay type.

AUSTIN P. EVANS,
Assistant Professor of History.

2. One of the most startling results of the New Examination has been to bring home to most of us the utter failure of our most careful efforts to eliminate our likes and dislikes for students from our estimate of their achievements, and to make this failure seem a very grave continuing chance for injustice and unreliability to persist in the old methods.

One of the irritation points of teaching is the uncertainty that fairness to the student has been accomplished. The New Examination relieves the teacher of the necessity for defending his position; an endless source of unpleasant bickering is eliminated.

Furthermore, the pedagogical value of examinations is appreciably increased. The feeling of the fruitlessness of mere markings is decreased in proportion as the feeling grows that students are stimulated to comprehensive and valuable review.

The premium on spotty cramming is removed. The certainty that examinations will be full as well as fair is a constant stimulation to consistent effort throughout the period of study covered by the examinations.

R. G. Tugwell,
Instructor in Economics.

3. The pedagogical value of these examinations lies first in the fact that a very wide field of information can be covered in the same amount of time, and still more in the fact that the student must make his ideas on definite questions clear and distinct. The only disadvantage from the pedagogical point of view is that the student receives no practice in expounding his views at length; but the quality of exposition made possible for the average student within an examination period is not high enough to make this consideration very serious. If examinations aim to provide a basis for grading and at the same time to discipline a student's judgment, this seems to be a good type of examination.

The experimental proof of the reliability of the examination is, of course, apparent in the correlations and other objective facts cited in the report.

Horace L. Friess,
Instructor in Philosophy.

4. I believe the testing of accomplishment by means of a paper made up of several fifteen or twenty-minute essay questions is fair neither to student nor examiner. If the questions require the marshaling of material from a range of sources, the vaguest generalities only can be stated in the time allowed. If they call for specific information, the range of topics will be too limited for an adequate estimate of the student's grasp of the whole subject. Marking is bound to be subjective, and the preservation of uniform standards throughout a long series of papers of the essay type most difficult for the examiner.

I believe the new type of examination, to be most successful, should include two parts, namely, (*a*) A two-hour examination of the new type; and (*b*) A one-hour paper consisting of one or two essays upon topics selected from a fairly large number of choices.

The chief points in favor of the New Examination seem to be:

(*a*) Complete objectivity in marking, making the marking of examiners identical, and insuring uniformity no matter how long or how numerous the marking sessions. Despite the scope of the paper, no longer time is required in marking it.

(*b*) Both specific information and ability to make judgments involving a thorough understanding of the subject matter of the course are tested.

(*c*) Thoroughness—information or judgment called for is specific in all cases, making vague generalizing impossible.

(*d*) Comprehensiveness—no student can complain that the examiner hit upon his weak points only. The entire range of the course is covered.

(*e*) Reliability—the above features make one more confident that a just estimate of the student's achievement is made than heretofore.

The second part of the examination is a necessary complement of the first. It emphasizes the synthetic aspect in essays upon carefully chosen topics involving the organization of material from a variety of sources more effectively than under the old system. The mere selection from the list is a test of the student in itself. The fewness and length of the essays invite more care on the part of the student than did those under the old system.

A. Gordon Dewey,
Instructor in Government.

5. As to my own standards of comparing students with each other I have never felt any great confidence or faith, and I welcomed this opportunity to get an objective measuring device. In its operation I could see no flaws either in its objectivity or in its ability to test the actual performance of the members of my class as students. It seemed to me much less subjective than my previous individual tests, although I have always attempted to hide identity by turning under names, etc., and having my wife check up by recorrecting cases that seemed doubtful.

The New type examination was accepted by my students with practically no exception as requiring more extended and searching preparation. Many have given me personal testimony to the effect that they feel they are more thoroughly tested and that they are obliged to work harder in preparation

in order to make a good showing. Furthermore, they like it and respect its results. The attitude assumed is that of an interesting and exciting game rather than two hours of drudgery. There are no post-mortem requests for change of grades, special dispensations, etc.

Many of us feel that in a broad and comprehensive survey course such as Contemporary Civilization there is a constant tendency to develop easy and facile repetition of generalizations transmitted through texts without the student's seeing the facts behind on which these generalizations depend. This new type examination has the pedagogical advantage that it discourages this sort of preparation and expression by the students. It is especially adapted to asking for judgments based on specific significant facts. In the old or essay-type examination this type of testing could be obtained only by selecting a few of the facts, and running the risk of being unfair, because of the inadequate sampling of the semester's work.

With the new type of examination with its adequate testing of fact-bound judgments, the essay type is removed from its double task and set free to test more adequately what it is best designed to test: comprehensive organization of relationships; expression of a general synthesis or philosophy of the entire field of study. Very interesting experiments in these essay questions might be attempted, which might develop more adequately than has before been done, the inventive power of students in seeing unconventional relationships and expressing a philosophy based upon the substantial background of a considerable period of study.

Robert D. Leigh,
Lecturer in Government.

6. The new form of multiple question objective examination gave a standard measure practically free from personal judgments. In addition it enabled the examiners to set questions on the whole content of the course. The students were pleased at the fairness of the examination, and, on the whole, preferred it to the essay form.

Experience with the handling of the test convinced me that it was much less of a mere fact collector than most of us feared. Judgments and inferences were required even in the *true-false* section. In spite of this fact I believe that a part

of the examination time, one-fourth at least, should be given to a comprehensive question on a subject matter requiring the weighing of evidence or the balancing of possibilities.

For small groups the labor of preparation and scoring is markedly greater than that required by the old method. Instructors who mark the tests themselves will be even more bored than when marking the essay type, and in addition will find themselves at a loss to weight the test and equate the score with letter-grades unless they have technically trained statistical guidance.

The examination is used most satisfactorily for a large class. It should be a long examination, more for the sake of testing the students than for increasing the accuracy over a shorter test of the same type. Clerical assistance should be furnished for scoring, and the derivation of the final score should be in the hands of psychological statisticians.

John J. Coss,
Associate Professor of Philosophy
and Director Summer Session, Columbia University.

NEW EXAMINATION METHOD ADOPTED FOR CONTEMPORARY CIVILIZATION

The above-described experiment was so convincing to the teaching staff and executive heads of the department of Contemporary Civilization in Columbia College that it was at once decided to continue the use of the new method for the final examinations. Accordingly, the final examination for the second semester of the course was closely modeled on the form of the examination described and illustrated above. This second new type examination, for C.C. A2, was given in May, 1921, in the same way as the examination for C.C. A1 had been given.

In general the results of this second trial of the new examination method confirmed the results reported above for the first trial. The distribution of scores for C.C. A2 was very similar to that obtained in January for C.C. A1,

and the correlation results are not different from those reported above.

The average correlation of the new examination scores is higher with Estimated Achievement than with Essay Examination marks. The range of the correlations with esti-

TABLE 45

Correlations of C.C. A2 New Examination Scores with Estimated Achievement and Hour Essay marks for each of thirteen classes. Average N = 35.

Instructor	A2 Estimated Achievement	A2 Hour Essay Mark
1	.67	.56
2	.76	.67
3	.59	.62
4	.87	.85
5	.71	.75
6	.94	.59
7	.74	.71
8	.70	.65
9	.54	.61
10	.59	.34
11		.83
12	.82	.84
13	.60	.67
Average r's	.714	.668

mated achievement is 0.34 to 0.88, and with essay marks 0.34 to 0.85. The higher correlations tend to go with the instructors who are older and more experienced as teachers.

The correlations in Tables 45 and 46 between the new examination and essay marks are higher than the reliability of the essay examination. This indicates spurious influences. It is quite probable that the instructors were influenced to a certain extent in making their C.C. A2 estimates by the grades which their students had received

in C.C. A1; and this influence also undoubtedly carried over to the marking of the essay papers. But this spurious influence is not large, and in any case, the cor-

TABLE 46

Correlations between C.C. A2 New Examination Scores and indicated variables for two groups of C.C. A2 students:
I. 208 men having the same instructor throughout the year;
II. 146 men who changed instructors at midyear.

	I	II
C.C. A2 Estimated Achievement Marks	.76	.77
C.C. A2 Hour Essay Marks	.64	.69
C.C. A1 Estimated Achievement Marks	.59	.59
C.C. A1 Hour Essay Marks	.59	.55
C.C. A1 New Examination Scores	.83	.88

relations tend to show that the new examination measures academic achievement as completely as the essay examination, and with greater reliability.

EFFECT ON INDEX OF COLLEGE SUCCESS

In a preceding chapter the hope was expressed that the use of objective examinations might give rise to more reliable indices of college success. Under the old system, as was shown in Chapter Six, our best index of college success has a reliability coefficient of only about 0.70. This means that a scholarship score of one year will predict that of the succeeding year with an accuracy no greater than that indicated by a correlation of 0.70. This low predictive value of our present index of college work may be explained in two ways: (1) It may be due to the imperfection and unreliability of our measures; that is, college success itself may be very constant from year to year, but our measures of it may be so imperfect

that *apparent* variations are indicated where none exist. (2) Or it may be due to the fact that college success itself is very inconstant and fluctuating from year to year. Certainly many complex elements enter into college success which are known to be variable if not capricious. Health, interests, distractions, griefs, worries, etc., often come and go in quite unpredictable successions.

It seems almost certain that these two explanations are complementary. We know that college success is incon-

TABLE 47

Correlations between C.C. A1 and C.C. A2 variables for the two groups of students described in Table 46.

	I	II
Estimated Achievement	.71	.65
Hour Essay	.64	.53
Final Mark (Ave. of New and Old Examinations)	.80	.77

stant, and we know that our measures of it are imperfect. The question is whether our measures *can* be improved in point of reliability, or whether they have already reached the limits of reliability set by the inconstancy of college success. The use of the objective and comprehensive New Examination in C.C. A1 and C.C. A2 has yielded some interesting data on this point. If college success is more constant than is indicated by a reliability coefficient of 0.70, then a better measure of it would give a reliability coefficient higher than 0.70. The correlations between first and second semester marks in Contemporary Civilization have been calculated for two groups of students with results as in Table 47.

The correlation between first and second semester estimated achievement marks is 0.71 for the group of stu-

dents having the same instructor both semesters, and 0.65 for the group that changed instructors at midyear. Corresponding correlations based on Hour Essay marks are 0.64 and 0.53. These represent the constancy of college success as measured by the traditional subjective devices. When each semester's marks are based partly on the new objective examination, the constancy of college success

TABLE 48

Correlations between C.C. A1 and C.C. A2 indicated variables for each of three classes. Average N = 30.

	Class I	Class II	Class III	Ave.
New Examination Scores...	.88	.75	.83	.82
Estimated Achievement	.34	.73	.76	.61
Hour Essay Marks	.49	.57	.70	.59

appears to be somewhat greater, the correlations for the two groups going up to 0.80 and 0.77. This indicates that with the traditional examination we have not reached the limits of reliability set by the admitted inconstancy of college success itself.

The data of Table 48 support this conclusion. Class I changed instructors at midyear; Class II was augmented at midyear by the addition of eleven students from another class; Class III had the same instructor throughout both semesters. The average correlation between first and second semester marks based wholly on the New Examination is 0.82; the average correlation for instructor's estimate of achievement is 0.61 and the Hour Essay marks 0.59. The conclusion seems clear that while college success is certainly inconstant from semester to semester and year to year, it is not as inconstant (in a single course) as it appears to be when measured by the

traditional methods. Moreover, it seems that in the new examination method definite promise is held out for the derivation of an index of college success which will be a much safer criterion whereby to judge the merits of admission methods as well as for the general administration of the college.

EXTENSION OF USE OF NEW EXAMINATION METHOD IN CONTEMPORARY CIVILIZATION

The comprehensiveness, reliability, and economy of the new method at once recommended its use in the hour quizzes that are given two or three times during each semester. The policy of the Dean of Columbia College, while not paternalistic, calls for continuous accounting for the academic status of every student, especially at the beginning of each session, and very especially for each beginning class. The kind of start that a Freshman makes very frequently describes his whole college career; and often a word of counsel from the Dean, based on exact information concerning a boy's academic status, may be far-reaching in its effect, especially if it comes early in the first semester in college. The "Four-Weeks Marks" are a very important element in carrying out this policy.

The difficulties of getting accurate marks on a common standard after only four weeks, particularly for the Freshman class, are obvious. The instructors have not become very well acquainted with their students in our typically large classes; and the evidences of a non-personal sort on the basis of which instructors may differentiate their students are very meager and insecure. The possibility of getting an objective rating of the Freshmen (all of whom are required to take Contemporary Civilization A1-A2) on a common standard and on the basis of a fairly large sampling of their first-year

work, was taken advantage of in October, 1921, when it was decided to use a True-False form of test in the derivation of the "Four-Weeks Marks" for the five hundred odd men in Contemporary Civilization. A series of eighty True-False questions was prepared by three of the instructors.

Each instructor took approximately one-third of the work to be covered in the examination and prepared a series of questions in typewritten form which were submitted for checking to the other two instructors. Practically nine-tenths of the questions submitted were accepted without alteration. In the committee, about eight or ten of the questions were revised or replaced by new ones.

The time taken in the preparation of the questions for the committee was about one and one-half hours each. The time taken in committee was about forty-five minutes. Each of the members of the committee had had considerable experience in making out this type of question and had made out questions covering the same field once or twice before.

In general, this specific committee procedure is recommended as a practicable and expeditious way to prepare these questions. It is important that each member of the committee have a specific field; that he remain close to the syllabus in preparing questions; that, if possible, he prepare in completed form his assigned portion of the questions on a basis of a schedule of number of questions assigned to each field covered in the examination, with one or two extra questions if convenient; that he prepare his questions on the typewriter *in triplicate;* and that these sets of questions be exchanged before the meeting of the committee.

The marks thus derived proved to be very accurate and reliable according to the best judgment of the instructors. For this reason the Dean found them very helpful in ad-

vising Freshmen in the critical period of the first two months in college.

The cost in terms of time and money of the New Type quiz was studied in the case of the mid-term examinations in December, 1921. The procedure was the same as in the case just described, and was reported to the Dean by the chairman of the committee in the following Memorandum:

The preparation of the examination, which consisted of eighty True-False questions, took about four hours on the part of one instructor, two of another, and one of another. The critical reading of the questions probably took an average of fifteen minutes from each of the other instructors. In accordance with directions from the Dean, clerks were employed at fifty cents an hour to do all the scoring, the total cost of this being $16.50, an average of less than four cents per student examined. The examination required forty minutes on the part of each examinee; in some classes, only twenty minutes were required, the rest of the hour being devoted to instruction.

The scoring, casting of the distribution, and transmutation of raw into letter-grade scores was done under the supervision of the writer, without any work or worry on the part of the instructing staff, their time being thus saved for services which could not be purchased for fifty cents an hour.

The reliability of the examination is estimated to be at least $r_{11} = 0.80$, and the significance, as attested by the agreement with instructors' estimates, equally satisfactory.

This might be contrasted with the situation if the essay type of examination had been used. The reliability would have been, if the reading had been as careful as during the experiment in 1920, $r_{11} = 0.67$, and the significance, therefore, not above 0.70. It was found that in 1920 the instructors were able to read not more than seven Hour Essay papers per hour with that degree of reliability. This would have required sixty-four hours of work on the part of the instructors, instead of only ten with the new form of examination here used. Including the cost of materials and printing, the new form of examination in this instance saved at least fifty hours of drudgery for a very busy staff of instructors, at a cost of not

over five cents per pupil examined, and with a reliability and significance impossible of achievement with the old type examination.

The time and cost of these objective examinations, after the brief experience of one year, have been reduced to very small figures. During the current year, it is conservatively estimated that at least 200 hours of bitter and almost profitless drudgery have been saved for a very busy staff by the fact that they have been rescued from the reading of about 2,000 one-hour essay examination papers through the use of the new examination method in the four one-hour quizzes that have been given in Contemporary Civilization. This freeing of the staff for more productive labors has, at the same time, been accompanied by an increase in the reliability and significance of the college grades in C.C. A1-A2.

The assignment of letter-grades on the basis of these objective quizzes has been in accordance with the principles set forth in the last chapter, and similar in procedure to the method described earlier in this chapter. The "passing line" was fixed by the consensus of opinion of the members of the staff. The following percentages are typical:

F	5%
D—, D, D+	10%
C—, C, C+	45%
B—, B, B+	35%
A—, A, A +	5%

This scheme is typical only for the whole Freshman class. The proportions receiving each letter-grade in individual classes vary enormously. In some sections eighty per cent of the students have received A or B; in others as high as forty per cent have received F or D. This suggests one of the greatest advantages of the objective type of examination: whereas with the subjective type of examination each section of twenty or thirty men

is marked on its own standard, every man in the course is marked on the same standard when the new examination method is used. This insures that, within the limits of the probable error of the new type scores, a letter-grade given in one section will be equivalent to the same letter-grade given in every other section.

The total cost of the *final* examinations in C.C. of the new type has been less than twenty cents per student examined, counting all the scoring as done by clerks at fifty cents per hour. The scoring could be done by cheaper labor, since it is perfectly mechanical. The printing cost has been about ten cents per blank, and the scoring a little less.

This may seem extravagant, but it must be remembered that in each final examination of the new type at least one hundred hours of reading of essays is spared the instructing staff, and their time and energy thus liberated for more economical employment. It must also be said that the cost of printing can be reduced by half after the current year, and that the cost of scoring will be reduced by one third, due to improvements suggested by a year's experience. The use, for example, of "+" and "0" instead of the "+" and "—" in the True-False test effects a saving of twenty per cent in the time required for scoring. The use of numbers in the Recognition form of test instead of underlinings of words saves at least a fifth of the time required for scoring that test. Even if the instructors score the new type examination, the time required is only one third of that required for scoring the essay examination.

The conclusion is that on every count the new examination is more economical in absolute terms, and more economical in view of the reliability and significance of the results obtained, than the traditional method of examining.

CHAPTER NINE

The New Type Examination in Physics

The experience of the Department of Physics can be introduced most satisfactorily by the quotation of a memorandum kindly furnished to the writer by Professor H. W. Farwell, whose experiments with New Type tests carried out over a series of years in collaboration with Professor Wendell constitute one of the most interesting and significant of the contributions in this field that have come to the attention of the writer. The following is a memorandum dated December 10, 1921:

For a number of years I have been experimenting with questions out of the stereotyped form, but not until the present year have I gone to any great extent over into the New Type of test. This year I have used the New Type test in every one of our weekly quizzes, usually in conjunction with a problem test such as we have been in the habit of giving. It seems to have been shown that the New Type is very useful in physics for the following reasons:

1. It supplements the problem test in a very desirable way.
2. It enables the quiz to cover more ground than the ordinary type covers.
3. It is more interesting and more stimulating to the average student.
4. It can be graded much more quickly and uniformly.
5. It can be made as difficult as one desires.
6. Used with the problem test it is more thorough than problems alone, or problems with the usual questions.

My experience would lead me to say that a test of the New Type to be satisfactory must be prepared with the greatest care, but that the results are worth the trouble.

Sample questions from the New Type tests used for the first five weekly quizzes in the first half of the semester are reproduced below. The first of the five proved to be

the least satisfactory one of the lot. The main apparent weakness was the failure on the part of the students to understand the directions; but a large defect, over and above the inadequacy of the directions, is the unreliability due to inadequate sampling of material and of performances, to complications of scoring which rendered objectivity almost impossible, and to the natural defects of the "matching" test.

SAMPLE QUESTIONS FROM FIVE NEW TYPE PHYSICS QUIZZES[1]

1. Physics Quiz

Among the principles so far considered the following may be noted:

A. Uniform motion.
B. Uniformly accelerated motion.
C. Statics of a particle.
D. Statics of a rigid body.
E. Conservation of energy.
F. Kinetics of a body.

The application of these principles involves, among others, the following concepts:

G. Power.
H. Work.
I. Friction.
J. Efficiency.
K. Kinetic energy.
L. Momentum.

Note below against the number of each problem the letter indicating the essential principles and concepts involved in the solution of the problem[2]

1.
2.
3.
4.
5.

Name.....

[1] It is important for the reader to remember that the questions reproduced are *samples* selected at random from the complete tests as given. The complete tests usually contained from two to four times the number of questions reproduced here. This is generally true of all New Type examinations illustrated by sample questions in this volume.

[2] In the quiz, the problems were on a second sheet. Since they are problems in physics of the usual problem type, they are omitted here.

2. Physics Quiz

Put the plus sign before each correct statement; the minus sign before each wrong one.

1. Rolling friction is less than sliding friction.
2. No machine can possess an efficiency greater than 100 per cent.
3. The mechanical advantage of a machine is independent of the work done.
4. If a machine has a high efficiency it must have a high M.A.
5. One pound force is the pull of the earth on the standard pound.
6. A body at the center of the earth would have zero weight.
7. An acceleration can be produced only by a force.
8. Weight in the C.G.S. system should be expressed in dynes.
9. A ball moving at the rate of 60 ft./sec. must have a constant velocity.
10. Impulse is the same as force.
11. A constant force is necessary to produce a constant velocity.

Name.....

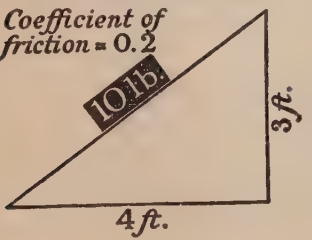
3. Physics Quiz

Fill the blank spaces with appropriate words.

1. The closer the stream lines in a steadily moving fluid, the greater the ..... of the fluid.
2. By Toricelli's theorem the velocity of efflux of a liquid is proportional to the .....
3. The energy of a steadily moving stream consists of ..... parts, one being due to ..... and .....
4. A substance which has the same properties in all directions at a given point is said to be .....
5. Strain is the ..... of an elastic body.
6. Hooke's law is exactly true only for a ..... substance.
7. Young's modulus refers only to the elasticity of .....
8. The operation of the hydraulic ram is ..... atmospheric pressure.

4. Physics Quiz

Put answers in the blank spaces. Numerical computations should be indicated but need not be carried out.



1. The component of the weight parallel to the plane is .....
   The component perpendicular to the plane is .....
   The friction force is .....
   The body (will) (will not) move down the plane. (Cross out the wrong statement.)

2. A certain jack screw has 5 threads to the inch and is operated by a lever 3 feet long. A force of 50 lbs. is needed to raise a load of 1 ton. The screw is 18 inches long and 2 inches in diameter.
   When the weight has been raised a distance of one inch, the applied force must have moved .....
   The work done by the machine is then .....
   The work done on the machine is .....
   The efficiency of the machine is .....
   The maximum M.A. of the machine is .....
   The actual M.A. of the machine is .....
3. To calculate the acceleration when given initial and final velocity and the time interval use the expression
   .....
   To calculate the acceleration when given initial and final velocity and the distance covered use the expression
   ....

Name.....

5. Physics Quiz

Underline the word or phrase which will make the statement most nearly correct.

1. Illuminating gas gradually fills a room by (osmosis) (diffusion) (effusion).
3. A (colloidal) (molecular) (alcohol) solution shows the Tyndall cone.
3. Surface tension is measured in dynes per (cm.^{2}) (cm.) (cm.^{3}).
4. The pressure inside a soap bubble is (less than) (greater than) (equal to) the pressure inside a drop of the same size.
5. Two bubbles in a cup of coffee tend to (separate) (come together) (stay as they are).
6. Jurin's law (does) (does not) apply to mercury in glass.
7. The English botanist Brown in (1827) (1897) (1742) on observing with the microscope very fine particles held in suspense in a mass of water, discovered they were in constant (regular) (irregular) motion, and the (larger) (smaller) the particle the more lively was the motion observed.
8. The kinetic theory of gases gives (objective proof) (no conclusion at all) (satisfactory confirmation) of molecular reality.

Name.....

The results obtained from these New Type weekly quizzes were so promising that a trial of the New Type questions on a larger scale seemed amply justified. Accordingly, Professor Farwell and his staff prepared a comprehensive one-hundred-statement True-False examination

to be given as an integral part of the final examination in the course Physics B1. These one hundred statements constituted Part I of the final examination and occupied the first hour of the three-hour final examination period. The questions were distributed over the various divisions of Physics (Heat, Light, Sound, etc.) proportionately to the time spent upon those topics. The two parts of the final examination are illustrated below.

Sample questions from the final examination in Physics B1 given in January, 1922. Part I included 100 True-False statements; Part II seven problems. One hour was allowed for Part I; two hours for Part II. All the examinees finished Part I; very few finished the problems within the time limit. The results of this examination, including both Old and New Type questions, were so satisfactory that New Type questions will henceforth be a regular part of all physics examinations.

PHYSICS B1—FINAL EXAMINATION, JANUARY, 1922

PART I

Mark with the sign + each statement which you know is correct, with the sign — those you know are wrong. Do *not* mark those which you do not know. PUT THESE MARKS ON THE *left hand* SIDE OF THE PAPER.

.. 1. Inertia is a property of all matter.
.. 2. There is only one resultant displacement which can be found when the component displacements are given.
.. 3. The change in velocity is the same thing as the acceleration.
.. 4. Two masses are equal if they balance each other when placed in the pans of an equal arm balance in a vacuum.
.. 5. The sum of the X— components of a set of forces whose resultant is zero is necessarily zero.
.. 6. Double weighing is unnecessary if the balance is known to give the same result, whether the weights are put in the right or left pan.
.. 7. Mechanical energy cannot be obtained from heat energy.
.. 8. The acceleration produced by a force depends on the time during which the force acts.
.. 9. The total kinetic energy of two bodies remains unchanged after an inelastic impact.
..10. Assuming that a high jumper could be carried to the moon and there make the same muscular effort as here, he would need to do better than eight feet to "make the team."

..11. A uniform wheel with constant angular motion about its axis which is at rest is in equilibrium.
..12. Simple harmonic motion may be defined as a motion in which the acceleration is always proportional to the displacement from a fixed point and oppositely directed.
..13. The amplitude of a S.H.M. is equal to the diameter of its reference circle.
..14. The maximum velocity of a body moving with S.H.M. depends upon the amplitude.
..15. The second's pendulum has the same length all over the world.
..16. All plumb lines are parallel.
..17. In a small body of still water it is not necessarily true that the pressure be the same at every point five feet below the surface.
..18. Pascal's principle states that a body submerged in a fluid is buoyed up by a force equal to the weight of the fluid displaced.
..19. The specific gravity of water at 4° C. is one, whatever the standard.
..20. The famous Magdeburg hemispheres were held together by the cohesion of their polished adjoining surfaces.
..21. The stream lines in a moving fluid are closer together where the velocity is greater.
..22. If a fluid is moving through a pipe of varying cross-section, the pressure will be least at points where the cross-section is least.
..23. Whenever a wire is stretched beyond its elastic limit, it breaks.
..24. A colloidal solution is a true solution.
..25. Surface tension increases with increase in temperature.
..26. Gases of different densities in contact will always diffuse.
..27. A brass rod in a vacuum cannot transmit sound.
..28. Two sounds each caused by a simple harmonic wave do not differ in quality.
..29. It is impossible to have standing waves with longitudinal wave motion.
..30. Nodes and loops exist only in standing waves.
..31. The overtones of an open organ pipe include all of those given by a closed pipe with the same fundamental frequency.
..32. The frequency of a tuning fork depends only on the amount of matter it contains.
..33. A body acted upon by two equal forces at right angles to each other cannot be in equilibrium.
..34. This test is merely a memory test.

Part II

Answer all questions

1. A uniform beam 26 feet long leans against a smooth wall, while its lower end rests on a smooth horizontal surface. A

horizontal cord fastened to the center of the beam and to the wall prevents the beam from slipping. The beam weighs 200 lbs. and its lower end is 10 feet from the wall. Find the tension in the cord.

2. A bullet of mass 2 grams is fired into a block of wood whose mass is 2,000 grams. The block moves .2 meters over a horizontal surface, the coefficient of friction being .5. Find the velocity of the bullet, the initial energy of the bullet, and energy of the block immediately after impact. From bullet in motion to block at rest what energy transformations take place?
3. (a) A hollow mass of zinc weighs in air 142 grams, and in water 100 grams. Find the volume of the cavity. Given density of zinc as 7.1 $\frac{\text{gr.}}{\text{cm.}^3}$ and of water 1 $\frac{\text{gr.}}{\text{cm.}^3}$. What would this object weigh in oil of specific gravity .75? (Neglect buoyancy of air.)

   (b) If this object were to be weighed in air of standard density with brass weights, show whether or not it is necessary to make the buoyancy correction in order to have the weight correct to .1 mg. (Density of brass 8.4 $\frac{\text{gr.}}{\text{cm.}^3}$ of air .001293 $\frac{\text{gr.}}{\text{cm.}^3}$
4. Combine two simple waves moving in the same direction whose frequencies are in the ratio of 2: 1, amplitudes 1: 2, and whose initial phase difference is 180°.
   Explain why this combination will or will not cause stationary waves.

Question paper must be surrendered with answer book

The experience gained in the use of various forms of questions in the New Type weekly quizzes showed that the True-False form was best adapted to the conditions and purposes of the Physics Department, and therefore only True-False questions were included in Part I of the final examination. These were formulated with great care and checked by several members of the department. Professor Farwell gives well-merited emphasis to the need for careful checking of the questions by at least two competent persons. The best procedure seems to be that of having at least two members of the teaching staff attempt to answer the questions in a *bona fide* manner as though they were actually taking the examination. In this way the am-

biguities and irrelevancies that creep into such questions, especially when the maker of the questions has had little practice in making out New Type questions, can be eliminated with practical completeness. Wherever possible this checking process should be adhered to even after the maker of the questions has had considerable experience. Part II

TABLE 49

Correlations for various types of tests given in the course Physics B during the first semester, 1921-1922, Columbia College.

		II	V	VIII	IX
I	Five 20-minute weekly quizzes of New Type given in first half of semester	.65			
II	Five 20-minute weekly quizzes of Problem Type given in first half of semester				
III	Mid-term grade			.66	.67
IV	Composite grade on New Type tests given weekly in second half		.63	.60	.60
V	Composite grade on Problem Type tests given weekly in second half				
VI	Grade on Oral Quiz in Laboratory		.27	.31	.35
VII	Grade on written reports on laboratory experiments		.34	.21	.37
VIII	New Type Final Examination, 100 True-False statements		.52		
IX	Problem Type Final Examination, 7 comprehensive problems		.60	.65	
X	(V plus IX)			.69	

differs from the usual problem type of final examination only in that no definitions or discussions are called for. It was thought that these aspects were amply covered by the one hundred True-False questions of Part I.

The data of Table 49 establish beyond question the validity of the New Type examination for physics. It seems clear that the New Type examination correlates with the

Problem test as highly as the latter correlates with itself. The New Type quizzes given in the first half of the semester correlate with the contemporaneously given Problem quizzes to the extent of 0.65; for the second half, the corresponding correlation between New and Old Type quizzes, is 0.63. The reliability of the Problem tests is not above 0.70, and probably not above 0.65. This is indicated by the fact that the Final Problem Examination gives a correlation with the composite grade of all the Problem quizzes given during the second half of only 0.60. The Final Problem Examination gives a correlation with the mid-term grade of 0.67, which is only 0.01 greater than that given by the New Type Final Examination with the mid-term grade. The mid-term grade was made up very carefully on the basis of all the information available on each student; this information was of several kinds:

a. Scores on seven New Type weekly quizzes.
b. Scores on seven Problem type weekly quizzes.
c. Oral Laboratory quiz marks.
d. Scores on Written reports of experiments.

The mid-term grade may therefore be considered as a very reliable index of the ability and achievement of the men in the first half of Physics B. The fact, then, that the New and Problem Final Examinations correlate almost equally well with the mid-term grade indicates that they are of about equal significance, with the balance in favor of the New Type with respect to reliability. The reliability of the New Type is not less than 0.80, and is estimated at about 0.90, by analogy with similar one hundred question True-False tests whose reliability coefficients have been calculated.

There is almost conclusive evidence, however, that the New Type examination is superior not only in reliability

but also in significance. The measure of its significance here used has been its correlation with the Problem examination. But if the Problem examination is not more reliable than 0.65, as was indicated above, its correlation with the New Type examination could not legitimately be much greater than 0.65. If the index of significance of the New Type Examination is limited by the reliability of the criterion, then the index ought to improve with any improvement in the reliability of the criterion. This is exactly what happens when we combine the scores on the Final Problem examination with the scores on the Problem quizzes given during the last half of the semester. The correlation of the New Type Final Examination with this combined score is 0.69. It is to be remembered that this combined score represents not less than six hours of examination-writing on the part of each of the students, and at least one hundred fifty hours of examination-reading on the part of the instructors; whereas the New Type Final Examination required one hour from each student, and probably not over twenty hours from all the instructors together. The conclusion is inescapable that with the New Type examination in Physics we get better results at less expense.

A remark may be added on the value of the Oral quiz and other laboratory grades as measures of achievement. It is very low. The oral quiz and the written reports on laboratory work are excellent and indispensable teaching devices; but their use as measuring devices is unwarranted by the experimental evidence here produced. If it is maintained that the laboratory marks are good indices of ability and achievement in Physics, then it must be maintained with equal fervor that the problem examinations are not good measures. This may be so, but it is contrary to the best opinion available, and contradicts all the ex-

perimental evidences available. Both cannot be right, and since the crown has been won by the Problem and the New Type examinations, the conclusion is that in deciding the final marks for a semester, the oral quiz and laboratory report marks should be weighted very lightly, if at all.

The distribution of the one hundred statements of the True-False examination over various major divisions of the course in Physics made it possible to ascertain in an objective manner whether there were any relatively weak or strong spots in the teaching of the material in the several major divisions. For example, let us consider a hypothetical comparison between the achievements in the divisions of Light and Sound, each of which is represented in the Final examination by twenty statements. If the average number of questions answered correctly on Light were seventeen, and on Sound six, an investigation would certainly not be amiss. Professor Farwell did not fail to notice and make use of this important pedagogical advantage of the comprehensive and objective examination. An analysis of the answers to the one hundred questions may reveal a systematic weakness in some one division of the subject matter which could be corrected by a slight change in the method of presenting and teaching it. Or it might reveal systematic weaknesses within divisions which could easily be corrected if their existence were known. Studies of this sort are already under way and only a little time is required for the production of decisive results. Of this and other pedagogical uses of the objective test more will be said in the following chapters.

CHAPTER TEN

The New Type Examination in Government

The experience of the Department of Government in Columbia College presents a notable example of the application of the results of the objective type of examination to the definition and carrying into execution of fundamental administrative policy. Members of the teaching staff in Government have long attempted to apply the same standards of achievement in all their classes, but, according to Mr. Macmahon, Head of the Department, "It is safe to say that the first accurate measurement of the situation has been made possible by the use of the objective and comprehensive New Type examination."

During the past year the course in elementary government was given in three branches of Columbia University, namely, Columbia College, Barnard College (for Women), and the Extension Department. Most of the instructors had classes both in Columbia College and in Extension. All the instructors used the same syllabus and the same method of instruction and recitation. The environmental factors affecting the learning of government in the three branches of Columbia University have been as nearly equal as they could possibly be made. Up to the present year the opinion has been current that in so far as classes in government were concerned the three branches of the university named were about equal in achievement. The records show that for some time past about equal percentages of failures in elementary government have occurred in the three branches.

During the first semester of the session 1921-22 several quizzes of the New Type were given in the classes in elementary government in all three branches. Each of these objective quizzes gave results which were consistently

contrary to the supposition of equality mentioned above. Large differences in achievement were discovered which had not been detected by the traditional examining and measuring devices. The results of the New Type final semester examination given to these classes in January, 1922, proved the existence of these large and hitherto undetected differences beyond the possibility of a reasonable doubt. This part of the final examination consisted of nearly two hundred True-False, Completion, and Recognition questions made and checked by all the members of the department. The comprehensiveness and objectivity of such an examination make its results strictly comparable for all groups. The following figures show the percentages of students getting each grade in Barnard College, Columbia College, and the Extension Department:

		F	D	C	B	A
New Type Government 1, Final Exam., January, 1922,	Barnard..			44%	33%	23%
	Columbia.	4%	8%	33%	42%	13%
	Extension.	24%	10%	39%	27%	0

Although these figures are conclusive enough, additional evidence has been furnished by Dr. Graper, who conducted a section of forty students in Columbia College and one of forty-two students in Extension. A carefully prepared examination consisting of eighty-five True-False questions was given to these two classes on the same day in March, 1922, with the following results:

	F	D	C	B	A
Graper's Columbia Section	7%	13%	20%	43%	17%
Graper's Extension Section	14%	29%	31%	17%	9%

The comparative achievements of these two sections, meeting in the same building the same number of hours per week, under the same instructor, using the same syllabus, the same readings, and the same method of instruction, demonstrate that the difference in effective scholarship between College and Extension students is not only large but genuine and systematic.

The evidence is equally strong that this large and significant difference would not have been detected by the old subjective methods of measurement. The average grades received by indicated sections on the different bases used in deriving the final semester marks in January, 1922, are as follows:

TABLE 50

	Instructor	New Type Examination Grade	Instructor's Estimated Grade	Hour Essay Grade
Columbia College:				
Section I	A	88	C	C+
Section II	B	70	C	C
Extension Department:				
Section I	C	68	C	C
Section II	D	62	B—	B
Section III	E	56	C	C
Barnard College:				
Section I	E	82	C+	B—

These figures show that sections which average the same grade on the subjective scales (Estimates and Essay marks) are revealed as widely different in average ability or achievement in Government by the objective test. The two Columbia College sections both receive the average grade of C according to the estimates of the instructors, but according to the objective examination, the average of Section I is 0.5 sigma above the average of Section II.

The three sections in Extension show the same discrepancies in more exaggerated degrees. Sections I and III both average C on the basis of the Estimates of the instructors and the grades on the Hour Essay Examination; but the average achievement as measured by the New Examination of Section I is 0.5 (sigma) above that of Section III. The instructor of Section II estimates his men as averaging B —, and gives them grades on the essay examination averaging a flat B; but as measured by the objective examination, his section is 0.25 (sigma) below Section I which is rated at C on both the Estimates and Essay examination grades!

A pedagogical application of the results of the objective examination has been reported to the writer by Dr. Graper. The class in American Municipal Government was given four one-hour quizzes during the semester, the first three being of the usual essay type, the last a New Type examination consisting of eighty-five True-False questions. The validity of this last quiz is indicated by its correlation of 0.80 with the average of all the other marks given to the students during the semester, including the three essay quizzes. The reliability of the sampling of the important facts, relations and judgments of the course afforded by the New Type quiz suggested the analysis of the responses to the individual questions with a view to determining the location of any weak spots in the presentation of the subject matter that might exist. Every question in the quiz related to an important idea or series of ideas which the instructor had especially emphasized, and theoretically, if the instructor had achieved his purpose of emphasizing these ideas about equally, each question should have been answered by approximately equal percentages of the students. The results of the analysis are presented in Table 51.

Questions 9, 23, and 43 were answered incorrectly by no students, while questions 30 and 49 were answered incorrectly by eighty per cent of the students. Table 51 shows clearly the difficult or pedagogically neglected points, and furnished a guide to instructor and student alike for profitable discussion, intelligent reading, and

TABLE 51

Number of students answering incorrectly	Designation of questions answered incorrectly
24	30 49
20	14
19	58
17	37 73
16	4 65
15	36 44 55
14	46 52 61
12	8 28 50
11	10 11 18 64
10	1 60 62 84 85
9	6 39 47 51 56 78
8	31 74 75
7	26 32 66 76
6	19 21 27 33 40 42 53 82
5	16 25 45 77
4	20 24 35 48 54 68 69
3	2 3 38 41 63 70 79
2	7 12 15 17 22 29 57 59 71 80
1	5 13 34 67 72 81 83
0	9 23 43

directed study. Dr. Graper has stated that further analysis of the results of the objective and comprehensive examinations will enable him to reorganize the course and redistribute the time, reading and discussion allotments in accordance with empirically determined needs.

It should be noted that one of the prime motives which led the Department of Government to experiment with the New Type examination was the desire to take advantage

of its direct pedagogical value in stimulating proper habits of study among students. Members of the department have long felt the need for checking the tendency of many students to content themselves with sweeping generalizations without any exact or comprehensive knowledge of the facts on which they are based. In the essay type examination it is very easy for the artful bluffer to deceive the most sagacious instructor by filling his discussions with empty generalizations. With the New Type comprehensive examination the students themselves admit that they are forced to learn the important details of factual foundations instead of merely memorizing certain formulas.

The procedure in making out the New Type government examinations was the same as that already described for the examinations in Contemporary Civilization and Physics. The first step was to assign to each division of the subject matter of the course its proportionate share of questions. The subject matter was then divided between the three members of the examination committee. Each member of the committee went through the material which had been assigned to him and selected a large number of important facts, relations or judgments to which the student should react in a definite way. Each of these selections was then put into that form of question which would elicit the right response from the understanding student and expose the ignorance or incapacity of the poor student in the surest and most economical manner. The extra time required for making out the New Type examination in government is more than compensated for in the saving effected by the objectivity in the scoring of the papers, and is in any case negligible in view of the ultimate results obtained. The New Type forms of questions are well adapted to the requirements of the courses in government, as seen from the following illustration:

Sample questions from the New Type final examination given to all classes in Elementary Government in Barnard College, Columbia College, and in the Extension Department of Columbia University in January, 1922. This examination consisted of a total of about 200 True-False, Completion, and Recognition items, and was allowed two hours of the final examination period. It was supplemented by a one-hour examination of the conventional essay type. The results convinced the members of the Government Department of the validity of the New Examination method, and henceforth the New Type questions will form a regular part of the examination machinery of the department.

MID-YEAR EXAMINATION, 1922

Government I

Name.....

Instructor's name.....

Part I

Directions for Part I. Read these statements and mark each one at the left of its number, with a plus sign if you think it is true, with a zero if you think it is false wholly or in part. Each statement marked correctly gives you a credit of one point; each incorrectly marked statement counts as a penalty against you, and is subtracted from your score; omitted statements do not count as penalties but reduce your total score.

Your score will be based upon plus and zero signs; don't waste time writing anything else.

First go through the list and mark all that you know for certain at once; then go back and study out the harder ones.

Do not guess. The chances are against you in guessing. A wrong response counts more heavily against your score than an omitted question.

A

.. 1. There are in the world to-day several republics that are not democracies, and a number of monarchies that are democracies.

.. 2. The Constitution of the United States established political democracy in this country.

.. 3. The states were responsible for the general abolition of property qualifications for voting.

.. 4. Residence qualifications for voting exist generally in the states of the United States.

.. 5. Labor organizations, because of their liberal tendencies, have generally favored free immigration into the United States, while organizations of capital, because of the radical tendencies of immigrants, have generally favored in the past a policy of rigid exclusion.

.. 6. Since only persons belonging to the Caucasian and African races may be naturalized in the United States no Mongolian can become an American citizen.

.. 7. A careful study of our early State papers will show that the founders of this republic were advocates of a two-party system.

.. 8. The Tammany assembly district leaders are not selected directly by the voters at the primary.

.. 9. Before 1880 political parties were not as a rule regulated by law. Since that time many legal restrictions have been placed upon them.

..10. The first great party controversy in the United States was between the Democrats and the Whigs.

..11. One of the laws passed by the Jeffersonian Republican party was the Sedition Act which did much to counteract the doctrines advocated by the leading Federalists who were responsible for the Virginia and Kentucky Resolutions.

..12. Although the Democratic platform of 1920 favored the league of nations as the surest practicable means of maintaining world peace, it did not oppose the acceptance of reservations making clearer and more specific the obligations of the United States to the league associates.

..13. The movement for manhood suffrage in the United States, as in Europe, was due almost entirely to the industrial revolution and the resulting growth of a numerically strong, disfranchised urban working population.

B

.. 1. With the outbreak of the American Revolution the newly-founded states set up for the first time legislative assemblies and began to gain experience in self-government.

.. 2. The state constitutions are in many cases amended by a referendum process but no such national referendum is now possible in the amendment of the national constitution.

.. 3. The first eight amendments are a guarantee of the individual against the national government only. We must look to the state constitutions alone for guarantees against state governmental action.

.. 4. The former distinction between constitutions and statutes is now losing force in the states because of the inclusion of much legislative matter in the state constitutions.

.. 5. The reaction to the defects of the Articles of Confederation caused a great centralization of powers in the constitution of 1787; other countries in copying features of our federal sys-

tem have been more jealous of central power and have reserved more functions to the states or provinces.

.. 6. The federal form of government requires for its successful operation the presidential system and, therefore, has been restricted entirely to the New World and the Pacific Islands.

.. 7. A significant economic clause in the Constitution is "Congress shall not make anything but gold and silver coin a tender in payment of debts."

.. 8. Under the 14th amendment Congress can punish hotelkeepers who refuse to admit negroes; Congress has failed to exercise this power on account of political opposition in the South.

.. 9. Although constitutional home rule affords adequate local protection against the state, it does not involve the federal principle relied upon by the states against the national government.

..10. The device of election of practically all public officials by the people directly, developed under the principles of Jacksonian democracy, was one of the first means by which the checks and balances system was diminished in scope.

..11. While the separation of powers is an important political theory it has not been recognized as a legal principle by the federal courts.

..12. The provisions of the U. S. Constitution and its amendments which guarantee private rights of person and property against state action are so broad that the bills of rights of state constitutions are superfluous.

..13. The bills of attainder issued by the English courts at the time of the Revolution were clearly prohibited in the American constitution.

C

.. 1. Bryce comes to the conclusion that second chambers are becoming less rather than more necessary.

.. 2. Congress often gerrymanders congressional districts.

.. 3. The legislature of North Dakota is interesting because of its recent experiments with the unicameral form.

.. 4. The President cannot veto items in appropriation bills; many state governors can.

.. 5. Since the two-party system is a sign of political maturity, there is a tendency, Bryce notes, for the number of parties in legislative bodies throughout the world to diminish.

.. 6. Woodrow Wilson's chief point, in the chapters on Congress in his Constitutional Government in the United States, was the fact that legislative leadership in Congress is concentrated.

.. 7. The Rules of the House of Representatives, once they have been adopted, can be changed by the Rules Committee without formality.

.. 8. The House of Representatives, as supposedly the more popular branch of the Congress, has allowed debate somewhat more freely than the Senate.

.. 9. Bryce indicates that direct legislation increases the chances of log-rolling between small groups.
..10. Bryce believes that direct legislation has been more successful in Switzerland than in the American states, although it has been relatively less used than in American states.
..11. Since deliberation and care are so much required state constitutional conventions are organized with the bicameral system.
..12. The latest developments in the forms of city government in the United States indicate a belief that, in view of the nature of the work of city governments, a representative body is not necessary.
..13. It is logically possible to defend the optional referendum although opposing the initiative.

Part II

Directions for Part II. Fill in the blank spaces so as to make sense and to convey accurate and significant information. Each unbroken line stands for one word, but you may use a phrase instead of a word if necessary. Read the question through to get its whole sense before filling in spaces.

A

1. There are in the United States two main types of ballot, the ..... type and the ..... type. In the former the various candidates are grouped according to the ..... they seek; in the latter according to the ..... The ..... type is the more common. In New York the ballot is of the .... type. Party leaders as a rule favor the ..... ballot because it encourages .... Reformers generally favor the ..... type since they think this type of ballot leads to ....
2. Two kinds of political officer may be distinguished, the ..... officer and the ..... officer. The former does the work of the ..... He is not paid for his work out of the .... but gets his reward in the form of ..... or ..... The latter performs the functions of the .... Of course one person may at the same time hold both kinds of office. An illustration of this is seen when a Tammany .... holds a ....

B

1. The economic groups or interests gaining most from a change of government in 1787 were ....., ..., and .....

2. The idea of a fundamental law superior to legislative enactments had its practical historical basis in colonial times in ...., its theoretical basis in the doctrine of ....
3. ..... is an example of a centralized, unitary state; ..... of a decentralized federal state; ..... of a more centralized federal state; ..... is an example of something that might be called a confederation.
4. In the United States the limitations on the national government are found in the .....; limitations on the state governments in the ....; and limitations on cities in the ...., ... and ....

C

1. The normal vote required to pass bills over the veto of the state governor is .....; but a practical reason why so few bills pass over his veto is the fact that a large majority of the bills are passed ....
2. The power of the speaker under American conditions lies chiefly in his power of ....., of ....., and of .....
3. The legislature in the United States cannot itself, without constitutional amendment, establish the initiative and referendum; this is a consequence of the doctrine of ...., which has as its corollary the rule that ....
4. The power of the speakership in the House of Representatives was notably enlarged by Speaker .....; it was curtailed in 1910 by a combination of ..... and .....; the Speaker was deprived of his power to name the .....; much of the Speaker's power has since been transferred to the .... and the ....

PART III

DIRECTIONS FOR PART III. Read these incomplete statements and their alternative endings and fill in at the left of the statement the number of the ending that will make the statement most sensible and true. It pays to make your choice in each case where you have a definite preference.

A

.. 1. The first Republican candidate for the presidency was: (1) Lincoln, (2) Douglas, (3) Webster, (4) Clay, (5) Fremont.

.. 2. The largest vote for a third party candidate for president between the years 1870 and 1908 was for the candidates of the: (1) Socialist-Labor party, (2) Prohibition party, (3) Populist party, (4) Greenback party.

.. 3. The South has remained almost solidly Democratic because: (1) the Southern people are democratic (2) most Democratic leaders are from the South, (3) of the race question in southern politics, (4) the Democratic party has consistently advocated the protection of southern agricultural products by a high tariff.

.. 4. In 1916 the Republican candidate for the presidency was: (1) Roosevelt, (2) Will Hays, (3) Lodge, (4) Hughes, (5) Taft.

.. 5. "We favor the granting of independence without unnecessary delay to the 10,500,000 inhabitants of the Philippine Islands." This was one of the planks of the 1920 platform of: (1) Democratic party, (2) Republican party.

.. 6. W. Wilson in his volume entitled Constitutional Government in the U. S. claims that: (1) parties are absolutely necessary in the United States to hold together the disconnected and dispersed governmental agencies, (2) political parties have tended to widen the breach between governmental agencies established by the Constitution, (3) in most countries parties are distinct authorities outside the formal government expressing their purposes through their own separate and peculiar organs and dictating what legislatures shall enact and executives shall address themselves to, (4) without political parties the voters of a country could be united more adequately upon national questions.

B

.. 1. The system of delegating power to the central government and reserving power to the states or provinces is found in the United States and: (1) Australia, (2) New Zealand, (3) Union of South Africa, (4) Canada.

.. 2. The system of reserving power in the central government and delegating power to the states or provinces is found in: (1) Switzerland, (2) Union of South Africa, (3) Canada, (4) Australia, (5) New Zealand.

.. 3. The New York-New Jersey plan of port development requires the consent of: (1) New York alone, (2) New York, New Jersey, and Congress, (3) New York and New Jersey, (4) the municipalities affected and the respective legislatures.

.. 4. Bryce holds that the best school of democracy is: (1) civic instruction in the grammar schools, (2) the practice of local self-government, (3) a government-owned and managed press.

.. 5. The doctrine of "freedom of contract," as a legal rule, is derived by the courts from the: (1) police power, (2) obligation of contracts clause, (3) due process of law clause, (4) commerce clause.

C

.. 1. The most usual device for the attempted control of the lobby is: (1) executive committee sessions, (2) registration of lobbyists, (3) encouragement of citizens associations to appear before legislatures, (4) abolition of the franking privilege.
.. 2. Differences between the House of Representatives and the Senate over the provisions of bills are usually adjusted by: (1) conference committees, (2) adoption of the Senate provisions, (3) re-reference to the respective committees of each House, (4) adjournment.
.. 3. Partly because of the location of the state capital, partly because of the structure of the committees, committee hearings have probably achieved their furthest development in the state of: (1) Indiana, (2) Illinois, (3) Massachusetts, (4) New York.
.. 4. The compulsory referendum has long been used in practically all states in connection with: (1) ordinary legislation, (2) constitutional amendments, (3) election laws, (4) acts vetoed by the governor.
.. 5. The compulsory referendum has been long used in states and cities in connection with: (1) regulation of business, (2) bond issues, (3) election laws, (4) the location of school buildings and parks.

In terms of significance and reliability, the examination just illustrated is, in consonance with the results described in the two preceding chapters, much superior to the old subjective examinations.

Its results correlate with the Estimates to the extent of 0.63, with the Essay grades 0.57, and with the average of the Estimates and Essay grades to the extent of 0.71. The apparent significance of the New Examination goes up as the reliability of the criterion is improved; its real significance is probably very much higher than 0.71. The correlation with the average of the Estimates and Essay grades of Part I alone is 0.60, of Part II alone 0.62, and of Part III alone 0.51. The Completion form of test seems the most efficient, thus confirming the result of the experiment with the classes in Contemporary Civilization A1, but again the difference is so small as to be negligible. The Recognition form of test (Part III), is the least

significant, but its correlation with the criterion is high enough to show that it adds considerably to the value of the New Examination.

The reliability of one-half of Part II alone, consisting of fifteen completion sentences, is 0.80; that of the whole examination is not less than 0.85, and is estimated at more than 0.90. The reliability of the Estimates and Essay grades may be surmised from the fact that they correlate with one another to the extent of 0.66. The only objection to this as a measure of their reliability is that it is very probably too high, since the Estimates were made *after* the essay papers were graded.

CHAPTER ELEVEN

The New Type Examination in Zoölogy

The Department of Zoölogy used the New Examination method for the final mid-year examination in General Zoölogy in January, 1922. The experience of this department is set forth in the following memorandum, for which I am indebted to Professor Calkins:

I have tried out the objective comprehensive form of examination with my two classes in general biology and am very much pleased with the results. For the first time in the twenty-five years that I have taught this subject I feel that I can look the mediocre student in the face and tell him with confidence that his knowledge of this subject is inadequate. The examination covered the whole course, which consisted of two lectures a week, an hour's quiz, and four hours' laboratory work. This was the schedule for the class which I shall designate Class 1, consisting of 110 students, 106 of whom took the examination. The second class, in Extension, which I shall call Class 2, had exactly the same course for the same period with the exception of the one hour quiz per week, which was omitted. The second class was composed of eighty-six students of whom eighty-four took the test.

The distribution of the scores for the two classes, with the letter-grades assigned, are shown in Table 52.

The distribution of the scores in the two classes verified the prophecy that the Class 1 men would show a better knowledge of the subject because of the weekly quiz than the men of Class 2. The relative scores of the two classes give almost mathematical proof of the value of the extra hour quiz.

As a time-saving method of examining, this method appeals to me very strongly. A great deal of time was taken in getting out these statements, but this could be done at leisure. It would be much simplified if after each lecture a set of statements connected with the subject discussed were drawn up. These could then be arranged for examination papers without

consuming much time. In scoring or in marking the examination books the time saved to the examiner is of the greatest importance. The scoring is largely mechanical and calls for little or no mental effort. The results are free from the disadvantage of the personal equation of different men marking the papers. Formerly for this examination we required not less than ten days to read the papers and grade the men. This year the grades were ready in forty-eight hours.

TABLE 52

New Examination Total Scores in Zoölogy	Class 1		Class 2	
—10- 0			1	
1- 9	1		1	
10-19	2			
20	3		3	F
30	3	F	11	D
40	9		10	
50	17	D	6	C—
60	9	C—	15	C
70	10		16	
80	12	C	7	C+
90	16	C+	9	B
100	9	B—	1	
110	10	B	2	A
120-129	2	A	2	
130-139	3			
N	106		84	
M	77		63	
Sigma	28		26	

The scoring might have been done even more rapidly had we not assigned different values to the different statements. A point of fundamental importance, ignorance of which was inexcusable, was valued at one, whereas if he failed to get a correct answer there is more discredit than for some subject of minor importance. Hence an incorrect mark for such a statement was -2 or -3, according to the statement. It was extremely interesting to note in the two classes the relative numbers of those who guessed. With the different weights

given to the different statements guessing invariably proved costly. Another means of obviating guessing employed in this examination was to bring out the same fundamental principle in at least two of the three parts of the examination or by different statements involving the same principle in Part 1 of the examination. This involves more time in scoring, but with rigorous examination by this method it would be possible to eliminate guessing almost entirely.

As to reliability, the comprehensive examination undoubtedly is a better test than the essay type. It is possible to cover the main points discussed in the entire course and thus give the student a better chance to show what he knows.

The one criticism that I would make is that, given a subject upon which to write, the student is called upon to collate his experience bearing upon that particular subject, which shows the power of co-ordination and of associating ideas. In this association he may omit one or more essential or important facts, in which case his mark suffers. It may be a subject on which he has only a hazy conception and the examination of the essay type may give him little opportunity to show what he has really gained. With the comprehensive type, on the other hand, we present a specific subject in the form of a statement which requires the co-ordination of his conceptions and calling for a decision in the light of his own experience in regard to the truth of a statement.

An unexpected advantage of this examination was shown by an unfortunate incident during the examination period. One man was suspected of copying. The evidence was circumstantial but was quite strong, and a comparison of the two books showed a remarkable relation. On the evidence of those two books alone one might easily have confirmed the suspicions of the proctors. I did not believe it myself, however, when the incident was reported to me immediately after the examination, so I called the young man to my office, told him of the suspicions, and told him that I did not believe them, and asked him if he would be willing to rewrite the examination in order to clear himself. He was delighted with the opportunity and sat right down without preparation of any kind and scored off his examination paper in twenty minutes. The correlation with his original paper was nearly perfect. I compared them with other papers of the same grade and found

the correlation quite as good as with that of the man he was suspected of having copied from. I then learned that the two men were intimate friends and had done their work together and prepared for the examination together, and made the same errors, but the man who was supposed to be copied made a score of 117 and the man who was suspected of copying scored ninety-seven. On the strength of this test I dismissed the complaint.

Another advantage which this method gives is that it shows what parts of the course as given should be strengthened, in what parts of the course the conceptions were hazy, and the need of emphasizing in the class room all such subjects. On the whole I am convinced of the utility and fairness of this type of examination for biology students and shall use it hereafter in all cases. A copy of the complete examination accompanies this memorandum.

MID-YEAR EXAMINATION, 1922

Zoölogy 1

PART I

In the following series of statements some are correct, some are not. Place a cross (X) in the space at the left of those that are correct and a circle (O) at the left of those that are incorrect. Study each statement carefully to determine exactly what is meant before marking. Confused or erased marks will count against you. Do not guess; if you are undecided leave the space blank. Regard each numbered paragraph as one statement.

.. 1. Destructive metabolism is due to the liberation of oxygen brought about through the agency of enzymes.
.. 2. Digestion of food is accomplished through the agency of hydrolyzing enzymes.
.. 3. Animal movements are due to the transformation of energy contained in proteins, carbohydrates, and fats.
.. 4. Bacteria are living enzymes.
.. 5. Yeasts, like other plants, absorb CO_2 and liberate O_2.
.. 6. Through the agency of bacteria alcohol is ultimately oxidized to form CO_2 and H_2O.
.. 7. Euglena, Paramecium, and Amoeba agree in possessing organs of excretion.
.. 8. Amoeba, Hydra, and Paramecium agree in possessing gastric vacuoles.
.. 9. The permanent gastrula of Hydra is formed by immigration of blastula cells.

..10. The absorption of digested food is accomplished by the same physical process in Hydra, earthworm, lobster, and tape-worm.
..11. There are seven longitudinal blood vessels in the earthworm, three of which are unpaired.
..12. The nervous systems of Hydra, earthworm, and lobster are differentiated into sensory and central nervous systems.
..13. A reflex action is brought about by the irritability of specialized cells.
..14. The prothallium of the fern and a medusa are analogous organisms.
..15. The sun's energy is captured by the green plant and used in the manufacture of proteins.
..16. Nitrates are useless to animals except as they come through plants.
..17. There is no loss of matter or energy in the cycle of living things.
..18. The body cavities of the earthworm and lobster are not homologous.
..19. The auditory organs of the lobster are situated at the bases of the antennules.
..20. In the tape-worm all organ systems are degenerated as a result of parasitism.
..21. Starch is carried about by the blood and distributed to the muscles where it is oxidized for muscular movement.
..22. Nephridia are aided by the chlorogogue cells in excretion.
..23. Natural immunity is the result of natural disease.
..24. Passive immunity is not acquired by treatment with diphtheria anti-toxin.
..25. Serology is a science dealing with antigens and anti-bodies.
..26. Individuals immune to tuberculosis do not require opsonins.
..27. Chromosomes are qualitatively different and retain their individuality.
..28. Both parents contribute equally to the offspring.
..29. The reduced number of chromosomes in egg and sperm of the same species is not always the same.
..30. Mendel's work in genetics was completed before the theory of evolution was advanced.
..31. A yellow (Y) wrinkled (W) pea crossed with a green (G) round (R) pea (yellow and round dominant) would yield on continued self-fertilization the following proportions of the different types of peas in the F_2 generation: 1GR: 3GW: 3YW: 9YR.
..32. A white eye female fly crossed with a red eye male would have progeny in the F_1 generation in which fifty per cent would be white eye males and fifty per cent would be red eye females.
..33. Natural selection may account for the perpetuation of species but it does not account for their origin.
..34. Inheritance of acquired characteristics is an established fact.
..35. An acquired characteristic if represented by a factor in the germ plasm, will be transmitted by inheritance.

..36. In synapsis the X chromosome is unmated or mates with a Y chromosome.
..37. In the fern reduction in the number of chromosomes takes place at the time of maturation of the germ cells.
..38. Mutations are sometimes produced by the loss or gain of chromosomes.
..39. In statement No. 32 above the result stated is possible only when the female is heterozygous for eye-color.
..40. In Mendelian inheritance of two pairs of alternative characters the proportion of pure dominants to pure recessives in the F_2 generation can be only 1 : 1.
..41.[1] Continued life of yeast cells without chlorophyll in Pasteur's fluid, is proof that animals can live without proteins.

Part II

Insert the appropriate word or words necessary to make a correct biological statement. Each blank indicates one word.

1. The manifestations of vitality are ....
.....
2. ..... ferments dissolved starch; ..... ferments emulsify fats.
3. Harmless living together without benefit to either, of different types of organisms is called .....
4. Supporting structures of the fern are called .....; of Hydra ..... and of the lobster .....
5. Conjugation in Paramecium is a process of ..... resulting in .....
6. Phagocytosis aids Hydra in .....; aids man in .....
7. The experiments with chlorophyll of .... indicate that ..... formation is a result of .....
8. Additional digestive surface of earthworm is given by the .....
9. Sperm cells of the earthworm are formed in ....; received during copulation in .... and fertilize eggs in the .....
10. In the lobster digestion takes place in the ....; absorption of food by the blood in the .....
11. Blood of lobster reaches heart from gills by way of the ....
12. A linking factor between two other chemical bodies is called .....
13. In Mendelian inheritance with one pair of alternative char-

[1] The Zoölogy examination as given contained 75 questions of this type. In Parts II and III also only one-half the questions in the complete examination are reproduced here.

acters the ratio in the F_2 generation is usually expressed as ....; of these one is .... one is .... and ....

PART III

Write in space at left of each question the number of the word or phrase which makes the most truthful statement.

.. 1. Specialization and differentiation in the higher animals and plants have been mainly in the interests of 1) Food getting; 2) Nutrition; 3) Reproduction.
.. 2. The external openings of the oviducts of the earthworm are in the 1) 12th; 2) 13th; 3) 14th; somite.
.. 3. Insects breathe by means of 1) Skin; 2) Lungs; 3) Gills; 4) Air tubes.
.. 4. Reproduction by sporulation is characteristic of 1) Yeast; 2) Chilomonas; 3) Amoeba.
.. 5. Reduction in number of chromosomes in the fern occurs in the 1) Archegonium; 2) Prothallium; 3) Sporangium.
.. 6. The nephridia of the lobster open to the outside on the 1) Antennules; 2) Antennæ.
.. 7. 1) Medusa; 2) Hydra; 3) Paramecium; is hermaphroditic.
.. 8. Immunity to typhoid fever is brought about by inoculation with 1) Anti-bodies; 2) Living bacteria; 3) Dead bacteria.
.. 9. Cleavage of the lobster egg is 1) superficial; 2) holoblastic.
.10. The mesoderm of the earthworm forms the 1) Typhlosole; 2) Blood vessels; 3) Central nervous system.
..11. The qualitative difference of chromosomes was proved experimentally by 1) Boveri; 2) Bateson; 3) Weismann; 4) Mendel.
..12. A white eye male fly crossed with a heterozygous red eye female produces male and female young in the proportion of 1) One red female: one red male: one white female: one white male; 2) fifty per cent red eye male and fifty per cent white eye female.
..13. The body hollows of both earthworm and lobster are correctly described as 1) Coelomic cavities; 2) Body cavities; 3) Lumena.

Two very interesting facts in this memorandum deserve special mention,—(1) the use of differential credits and penalties, and (2) the application of different standards to two classes in the same subject matter.

(1) The use of differential credits, which permits the weighting of questions according to their importance, is entirely justifiable, but to penalize wrong responses heav-

ily and irregularly tends to introduce new and complex chance influences. Guessing is adequately checkmated in a True-False test of over seventy questions by the Right-minus-Wrong method of scoring. When a series of questions is scored after various patterns, such as

1 Right minus 1 Wrong
3 Right minus 0 Wrong
1 Right minus 3 Wrong
1 Right minus 2 Wrong

very complex and probably large chance influences come into play, causing many slight and a few serious dislocations. Moreover, these variable patterns of credits seem inadvisable for various reasons:

(*a*) It is very difficult to weight individual questions in a list of seventy-five with any assurance of fairness, the only means of finding the difficulty of questions being the method used by the Department of Government reported above;

(*b*) In a list of seventy-five questions, the units are many and small, and therefore the differences in difficulty are probably distributed in a smooth continuous series, so that by the Right-minus-Wrong method of scoring the questions are very probably automatically weighted in the best possible way;

(*c*) Differential patterns increase the labor of scoring, increase the chances for mechanical errors in addition to the introduction of the chance elements mentioned above, are subjective in effect, and even under the best possible conditions probably add very little to the classifying efficiency of a seventy-five-question True-False test in which the final marks are based upon relative position in the distribution.

(2) The application of different standards to these two classes in general Zoölogy illustrates the principle of meas-

urement described in the summary of Chapter Seven. A measuring scale may give data upon which policies may be wisely framed, but it does not in itself create or dictate policy. The conclusion that the superiority of class 1 over class 2 ($M_1 = +.5$ sigma$_2$) is due mainly to the extra hour of instruction received each week by class 1 is very reasonable; but strictly speaking, this conclusion depends upon the assumption of equality in the average ability of the two classes. Professor Calkins has good reasons, not stated in the memorandum above, for assuming that the classes are of equal or very nearly equal ability. This is not the question which interests us at this point. Granting that class 2 is equal in ability to class 1 and that its average inferiority of 0.5 sigma in the examination is due entirely to the absence of the extra quiz-hour each week, does not this very hypothesis involve the admission that the students of class 2 have gained less and know less Zoölogy than those of class 1? Are we to give an F grade man a D, or a D grade man a C, merely because we know that if he had had another hour of class work per week he *could have actually* earned the higher grade? Carried to its logical extremity, this policy would not stand criticism. But the important point here is not whether this is good or bad policy, but that it *is policy*, and not an inherent element in the measuring device or examination used. If it seems desirable to report two students as of grade B in Zoölogy whose real grades are C and B, the New Examination method enables you to do that with great accuracy; but the New Examination method neither approves nor disapproves the policy. It will be recalled that the opposite policy was adopted by the Department of Government in which every student was marked on the same objective standard. In so far as the New Examination is accurate, a grade of B in Gov-

ernment is equal to every other grade of B in Government, whether in Extension or regular college classes. In Government the basis of marking is defined as "Knowledge of, and ability to think in, the material of the course." It is a basis of absolute *achievement* in *Government*, not estimated *possible* achievement. The basis of marking in Zoölogy, as described in the memorandum above, includes moral and equitable considerations suggestive of the tempering of justice with mercy. This is a human and defensible policy; but unless the basis of marks, and the policy of controlling their administration, are clearly defined, the grades thus derived may easily play havoc with every agency of education that interprets all identically-named grades as indicative of equal achievement in the specified subject matter.

The reliability of the Zoölogy New Examination reproduced above is 0.75. The intercorrelations of the Parts average 0.65.

CHAPTER TWELVE

The New Type Examination in Economics, Philosophy, Greek Art, History, English and Civil Engineering

The purpose of this chapter is to illustrate the applicability of New Type questions to other subject matters taught in colleges and high schools. The method of exposition will be that of showing sample questions from examinations that have actually been used, with only brief notes on the general results obtained from them.

ECONOMICS

The attitude of the Department of Economics after the first trial of the New Examination method is as favorable to the New Examination as that of the other departments whose experiments have been described in the preceding chapters. According to a statement of Professor E. E. Agger, Head of the Department, the results of the New Examination "greatly surpassed expectations," and led to the immediate adoption of the policy of making the New Type examination a regular part of the final examinations in economics. The New Type final examination given to 259 students in elementary economics in January, 1922, is adequately illustrated by the following sample questions:

MID-YEAR EXAMINATION, 1922[1]

Economics 1

PART I

1. Part I and Part II are to be answered in two hours.
2. In Part I, where you think a statement is true as it stands

[1] Only one-half the questions in the examination as given are reproduced here.

put a plus sign (+) at the left of the question. If any part of the question or all of it seems to you to be false, put a minus sign (−) at the left of the question.

3. As the final mark for Part I is obtained by subtracting from the total score made by adding the correct signs, the total of incorrect signs, it is advisable to leave those questions unmarked concerning which you are in doubt. In short, DO NOT GUESS.

PART I

.. 1. A good workable definition of "economics" is, the study of man in his efforts to get and to use a living.
.. 2. It should be understood that economics is a pure, not an applied science.
.. 3. Business always has as its chief aim the production of goods; the acquisitive activities are secondary.
.. 4. The social point of view need not be emphasized in economics because of the high degree of individualism of effort in our competitive system.
.. 5. The divisions of subject matter in economics are for the sake of convenience, not because of natural law which separates one part of the science from the others.
.. 6. The study of production includes transportation.
.. 7. Consumption, to distinguish it from distribution, should be understood as a process of rearranging utilities for the sake of greater convenience.
.. 8. Of most social sciences it is true to say that individual prejudice is a danger to clear thinking. This is especially true of economics because of our human inclination to be influenced by economic interest.
.. 9. The chief objects of economics as a study might be said to be:
(1) to explain how the industrial system works;
(2) to criticize the present system and to devise improvements.
..10. The need for a science of economics arises because of man's power to modify his environment by conscious effort.
..11. The essential features of private property have been the same since the time of primitive man.
..12. Among the early herdsmen, there was private property in herds and slaves but not in land.
..13. When the agricultural succeeded the pastoral stage of civilization, private ownership of land was at once instituted.
..14. Private property, according to the medieval philosophers, was justified as a natural right.
..15. Mutual aid and competition are incompatible with each other.
..16. Competition among animals is only for consumption; among men it often takes the form of production.
..17. Legal monopolies, which were once so common, have not entirely disappeared, even in the United States.
..18. Slavery flourished in the South rather than the North because there was less good land in the South.

..19. The disappearance of free land secures us against the reappearance of slavery.
..20. True economic liberty, according to Seligman, will result only when the government ceases to meddle with industry.
..21. A government bond is individual wealth but not social wealth.
..22. Since capital consists of tangible wealth, it is an error to measure it in dollars.
..23. Demand and supply are the determinants of prices; there is little control exercised by prices upon demand and supply.
..24. Price is not only a measure of value, but it is also a means by which the rewards are distributed in society for varying services rendered.
..25. The reason why *free goods* have no utility is that the supply is not limited.
..26. Division of labor in industry has greatly delayed the operation of the Malthusian doctrine of population.
..27. The insurance agent cannot be called a producer of wealth.
..28. The availability of a labor supply is both a cause and a result of localization of industry.
..29. Machine production has reduced the tendency to "bargain" and "haggle" in retail trade.
..30. Modern factory production has inhibited the exercise of the "creative impulse in industry."
..31. A function of the middleman is to encourage production where it is cheap and to discourage it where it is dear, and so to benefit society.
..32. Minute division of labor cannot exist without the use of complicated machinery.
..33. Circulating capital is that form of capital, which, like money, passes freely from hand to hand.
..34. The imposition of a heavy tax will very often decrease the amount of land in the same way that it decreases capital.
..35. The value of fertile land does not depend on distance from the market.
..36. The law of diminishing returns applies to urban land as well as to agricultural.
..37. Differentiation of function ultimately makes possible a higher general standard of living.
..38. The Malthusian theory states that the lack of means of subsistence is the proper check on overpopulation.
..39. The loss of capital due to preventable waste in the United States has been brought to a minimum.
..40. Immigration of cheap labor into this country has assisted the development of division of labor.
..41. The accumulation of capital is due entirely to the efforts of labor.
..42. In the last analysis all capital may be said to be due to saving.
..43. Anticipation of demand in production is necessary because of the desire of a few middlemen to speculate.
..44. The diffusion of risk aids industry and production.

..45. The order of a business cycle is depression, boom, crisis, depression.
..46. A depression is accompanied by "tight money."
..47. There is no such thing as general overproduction and when the term is used it merely signifies that consumers will not take the goods at the price asked.
..48. The more specialized the productive processes of the country become, the more highly organized society will be and these conditions will tend to prevent crises.
..49. Fluctuating price levels make profiteering possible.
..50. Government enterprise during a period of depression would be unwise from an economic standpoint.
..51. Competition tends to prevent selling dear and buying cheap. It is to the interest of competitors to extinguish competition within the group.
..52. Agricultural coöperation on the Continent is common and successful but the Rochdale Stores in England have been failures.
..53. Fully eighty per cent of the production in the United States is carried on by corporations, which indicates that combinations occupy most of the field and that the small concern is rapidly passing out of existence.
..54. Agreements in restraint of trade have been illegal and non-enforceable under the common law in the United States; the same applies to Germany.
..55. Consumption means, in economics, the use of goods in further production.
..56. The larger one's supply of a commodity, the smaller in general will be the importance one attaches to the possession of any one unit of the supply.
..57. The marginal utility of water will depend on the urgency of the less-important want of it.
..58. Marginal Utility enables one to ascertain the total utility of one's whole stock of the commodity.
..59. The margin of consumption is the same for different persons.
..60. Excessive expenditure is socially beneficial because it gives employment to labor.
..61. Harmful consumption usually tends to diminish the sum total of the durable satisfactions that the consumer gets out of life.
..62. Our wants are often imposed upon us by the force of imitation and by social standards.
..63. The standard of living of a class is the product of the ideals and resources of its members, and, in turn, modifies their ideals.
..64. In the determination of the standard of living income plays a negligible part.
..65. The ideal standard of living demands the satisfaction of reasonable wants of both body and intellect.
..66. Demand, as the term is used in economics, is synonymous with desire.
..67. Popular opposition to monopolies is a product of the Industrial Revolution.

..68. Railroads should be controlled through the restoration of competition.
..69. A monopoly without complete control over prices cannot be considered a real monopoly.
..70. It is useless to tax a monopoly because the tax is simply shifted to the consumer.
..71. The outcry against monopolies is reasonable because in the long run monopolies can influence the fundamental factors in wealth distribution.
..72. The monopoly problem cannot be solved by complete destruction.
..73. The sole purpose of combination is to gain monopoly profits by extortionate prices.
..74. Monopoly is not solely the outcome of efficiency.
..75. Conditions of cost may be such as to make monopoly price no higher than competitive price.

PART II

In Part II, select the alternative which you think is correct, and write the letter denoting this alternative at the left of the question. You are not penalized for errors in this part, unless your choice is obviously preposterous, so that it pays to take a chance, even where you are not quite sure.

.. 1. The chief subject matter of economics is: (a) the wealth creating activities of man, (b) the general laws of making wealth, (c) the uses and abuses of wealth, (d) the creating and using of goods and services.
.. 2. The right of property is to be justified, if at all, because it: (a) increases the legal liberty of all, (b) increases the social welfare, (c) gives to each what he has produced, (d) promotes economic equality.
.. 3. One would expect the highest development of division of labor in: (a) truck farming, (b) automobile repairing, (c) interior decorating, (d) carpentering, (e) shoe manufacturing.
.. 4. The bearer of risk in corporate enterprise is: (a) bond holder, (b) promoter, (c) share holder, (d) member of the board of directors, (e) manager of corporation.
.. 5. A holding company is: (a) a business union in which each unit coöperates with each other and shares profits jointly, (b) a company having under its control sufficient securities of certain subsidiaries to dictate their policy in at least one respect, (c) an amalgamation of two or more business units into one concern under a new single charter.
.. 6. During an incipient panic the policy of bankers should be: (a) to lend freely, (b) to restrict loans, (c) to refuse all loans.
.. 7. Luxury consists in the use of: (a) pernicious goods, (b) unnecessary goods, (c) expensive goods.

.. 8. In the early days in England monopoly rights were granted: (a) by the crown, (b) by parliament, (c) by the courts.
9. An example of relatively complete monopoly is: (a) the telephone company in New York City, (b) the Brotherhood of Railway trainmen, (c) the Sheffield Farms Milk Co.
..10. Some monopolists have class prices for their products. From the social point of view this is: (a) good, (b) bad, (c) good or bad according to the nature of the product or the other relevant circumstance, (d) a matter of complete indifference.

The distribution of New Examination scores, with letter-grades assigned, is shown in Table 53.

TABLE 53

New Examination Scores in Economics 1-2	Frequency	Letter-Grades
20-29	1	
30	4	
40	5	F
50	15	D
60	28	C−
70	34	C
80	40	C+
90	46	B−
100	38	B
110	27	B+
120	15	A−
130-139	6	A

As in the other subject matters, the New Examination in Economics correlates with the criterion as highly as the unreliability of the criterion allows. The reliability of the criterion in this case is 0.45, as measured by the correlation between the Estimates and Essay Examination grades. The correlation of the new examination scores with the Estimates is 0.34, with the Essay grades 0.37, and with the average of the Estimates and Essay grades 0.45. The correlation between the two parts of the New

Examination in Economics is 0.48; between random halves of Part I it is 0.71; and between random halves of the whole new examination the correlation is 0.73.

PHILOSOPHY

The New Type examination in the final mid-year examination in the course Introduction to Philosophy in January, 1922, consisted wholly of True-False questions. In addition to the New Type examination, illustrated below, several searching questions of the Essay type were included in the total examination. The distribution of New Examination scores for the 229 students who took the examination turned out to be a fairly smooth symmetrical curve with the Mean at thirty-seven and a Standard Deviation of eighteen points. The effective range was from eight to seventy-five points.

The correlation of the New Examination with Instructor's Estimates was 0.66, and with the Essay grades 0.59. This is in consonance with the rule that the New Examination correlates with the criterion up to the limit of the reliability of the criterion. The correlation of random halves of the New Examination in Philosophy I, consisting of 107 True-False questions, is 0.75.

MID-YEAR EXAMINATION, 1922[1]

Philosophy 1

The following statements are either true or false. If you think a statement true put a + (plus) sign in front of it; if false put a − (minus) sign. If any part of the statement is false the whole statement is considered false.

If there are statements which you consider debatable put + or − as the argument for its truth or falseness seems to you better. Put a circle around such a + or a −, but around these only. Such debatable questions are not those in which any part is clearly false.

[1] Only one-third of the questions of the complete examination as given are reproduced here.

Read through the entire examination and mark the statements of which you are sure. Then reread and consider the answers to those which appear more difficult.

A correct judgment counts 1, an incorrect —1, and an unanswered statement counts nothing. In general, it is better not to guess.

For this part of the examination one hour and fifteen minutes is allowed.

.. 1. No matter what inequality of wealth or position men suffer from, their fundamental equality of intelligence remains constant.

.. 2. The quick decisions and ready attention of the impulsive man render him a steady and reliable leader into new and profitable fields.

.. 3. In some cases of reflection the probability and desirability of a predicted outcome furnish the reason for the decision reached.

.. 4. By knowing the tests of good scientific thinking, we shall be able to make our own thinking of that type.

.. 5. The discovery of new mathematical truths, like that of all other kinds of knowledge, occurs by means of the imaginative elaboration of new suggestions, and differs materially from the deductive proof to be found in books on mathematical propositions.

.. 6. Most modern mathematicians maintain that neither intuition nor experience forms a satisfactory guarantee of the validity of geometrical axioms.

.. 7. The progress of science in the last three centuries has borne out Descartes' prophecy that mathematics is a key to the great cipher of Nature.

.. 8. In making his theorem more general Pythagoras reduced the number of initial assumptions with which he had to start.

.. 9. Ptolemy's interest was in describing the real universe, rather than the mere achievement of a harmonious system of calculations.

..10. Copernicus was forced to assume epicycles to account for exceptions in his system.

..11. Simpler hypotheses are preferable, since nature has been found to take the simplest course when possible.

..12. The discovery of Uranus involved the deductive application of hypotheses.

..13. Where the 18th century scientist thought heat was a fluid, we know to-day it is nothing but the energy of motion, and there is really no "hotness" at all.

..14. Rumford's ice experiment destroyed men's belief in the caloric theory of heat; it remained for Davy to establish definitely the molecular theory.

..15. If the ultimately simple elements out of which material bodies are formed are alone real, then the real world is the world of molecules.

..16. Darwin was not primarily interested in the origin of man when he began his investigations.
..17. The doctrine of uniformitarianism states that the past history of the earth has been marked by a regular series of universal catastrophes.
..18. The theory of evolution was first propounded by Charles Darwin.
..19. Continuity of variation in living forms makes their division into rigid forms an artificial matter.
..20. The theory of evolution attempts to explain the past by means of factors which are observed to operate in the present.
..21. The invention of the microscope brought to light facts which helped to support the erroneous theory of spontaneous generation.
..22. The method of agreement is easily applicable to any actual situation.
..23. The possibility of scientific knowledge rests on the assumption of a certain uniformity in nature.
..24. Since we have succeeded in isolating the typhoid bacillus, and know how to secure it from the blood of an infected person, it is now always easy to identify a case of typhoid fever.
..25. Since the aim of the scientist is to discover universal laws, he does not take into account the peculiarities which distinguish individual objects from one another.
..26. The tendency to confuse our feelings toward an object with the object itself is one of the chief obstacles to a scientific understanding of the world in which we live.
..27. The arrogance of the higher critics is evident in the very adjective they apply to themselves—"higher."
..28. The Pentateuch itself contains evidence that it came from the pens of at least four distinct writers.
..29. The three main supports of the usual theological view of the Bible are tradition, authority, and *a priori* reasoning.
..30. The second Ritchie Case did not deal with the same act of the legislature with which the first case did, but it involved the same general principles.
..31. The statistical facts presented to the judges in the second Ritchie Case convincingly proved the constitutionality of the act.
..32. Under the "police power" the constitution enumerates various types of legislation, such as the regulation of sanitary conditions in factories, safety appliances, etc., which are to be undertaken "to promote the health, comfort, safety, and welfare of society."
..33. There is an obvious and urgent need for judges who are better disciplined in the logic of constitutional law, who will be able to deduce their judgments with more scientific accuracy from the constitution, and who are therefore less subject to the influence of changing facts and party politics.
..34. A correlation of —1 indicates absence of causal relationship.

..35. Statistical measurement eliminates the personal factor in interpreting social data as to causes and values.
..36. In a normal curve of distribution, the greatest number of cases occurs near the average.
..37. Definitions of social ideals require more observation and classification of data than do those of geometry.
..38. Correlation is a development of the principle of concomitant variations.

The dividing line between C and D was decided by the teaching staff in the course, and was set at twenty-two, about one S.D. below the Mean. The assignment of letter-grades was as follows:

New Examination Scores in Philosophy I	Letter-Grades Assigned	Frequency
0-16	F	21
17-22	D	19
23-28	C−	24
29-36	C	40
37-42	C+	22
43-48	B−	26
49-56	B	26
57-60	B+	19
61-64	A−	12
65-70	A	16
71-78	A+	5

The variations in subjective standards are brought out by the use of the objective examination in Philosophy I just as in the other subject matters mentioned above. The averages of three types of grades for each of four instructors' classes follow:

	New Exam.	Estimates	Essay
Instructor I	39	C+	B−
Instructor II	37	B−	C+
Instructor III	36	C+	C+
Instructor IV	35	B−	B−

The class with the lowest average score on the New Type examination was estimated by its instructor at a higher grade than the estimated average grade of the section which received the highest average score on the New Examination. Section I, by the New Examination, stands about one-fifth of an S.D. above Section IV; but by the estimates of the instructors, the positions of the two sections are reversed. The fact that different standards are used by the instructors is also brought out by the disagreements between the average of their Estimates and of their Essay grades. A, B, C, and D mean different degrees of merit to different instructors, and to the same instructor at different times.

GREEK ART

Professor C. H. Young, Executive Officer of the Department of Greek and Latin, tried the New Type examination in a fortnightly quiz with his class in Greek Art in Barnard College, and reached certain very definite conclusions at once. The first was that the New Examination method has the tremendous advantage over the Old method of enabling the examiner to touch upon practically every part of the ground that has been covered in the course, instead of merely selecting a few high points here and there. The second was that for a class of considerable size (in the present case fifty-six), the New Type examination effects a distinct saving in time. The preparation of the paper did not demand more than twice the amount of time needed to prepare a paper of the Old Type, and the marking of the papers was accomplished in not more than one-eighth of the time usually required for marking the Old Type papers. The third conclusion was that the results were more accurate and just, not only because of the

objectivity of the scoring and the elimination of the personal equation, but also because of the better sampling of the materials of the course and of the performances of the students.

In order to compare the New and Old methods directly, Professor Young gave an hour essay quiz to the same class about a month after the New Type quiz had been given. He read the essay examinations very carefully and graded them without being influenced by the scores made previously on the New Type quiz. The correlation between the two sets of data was found to be 0.70. For the sake of illustrating the adaptability of New Type questions to such courses as Classical Civilization or Greek Art, the New Type quiz which gave this high correlation with the best independent criterion available is reproduced in full:

GREEK ARCHITECTURE QUIZ

I

1. Sphenoidal masonry was much used by the Greeks.
2. The Greeks used sundried brick, because they did not know how to make burnt brick.
3. Columns of inferior stone were coated with stucco.
4. Tessellated mosaic was made of natural pebbles.
5. The best Greek marble is found in Paros and on Mt. Pentelicus.
6. Stones of the same course were bonded with cramps.
7. Dowels were more commonly made of bronze than iron.
8. The Greek temple may have been derived from the Mycenæan megaron.
9. Columns were usually monolithic.
10. The flutes of Doric columns usually meet in points.
11. The Doric column has a base different from the Ionic.
12. Dentils occur regularly in the Doric order.
13. Mutules are never found in an Ionic building.
14. Metopes sometimes occur in the Corinthian entablature.
15. Entasis is more pronounced in Doric than Ionic columns.
16. Roof tiles were made of marble or of terracotta.
17. Horizontal and raking cornices are identical in shape in the Doric order.
18. An anta is an engaged pilaster.

19. An antefix may serve a practical as well as an ornamental purpose, whereas an acroterion is purely ornamental.
20. Consoles are found only in the Corinthian order.

II

1. The Doric order is more ..... than the Ionic, the latter less ..... than the Corinthian.
2. Blocks of soft stone were hoisted with ....., but marble blocks with ..... or .....
3. Anathynosis occurs in ..... joints, because of the lack of ..... at such joints.
4. Euthynteria is the upper course of the ....., whereas the upper step is the .....
5. The end and side views of a Doric capital are ....., and it is, therefore, ... to the Ionic capital for a corner column.
6. The most elaborate architectural ornament used by the Greeks is the ....., composed of alternating ..... and .....

III

1. Polygonal masonry is more enduring than tetragonal because: (1) the stones are usually larger, (2) the joints are irregular, (3) it is more carefully constructed.
2. The normal number of flutes in a Doric column was: 12, 16, 20, 24.
3. The Greeks as a rule bonded stones: (1) by gravity, (2) with mortar, (3) with cramps and dowels.
4. Pryholes were used to adjust a stone: (1) laterally, (2) vertically, (3) longitudinally.
5. The Corinthian order is distinguished from the Ionic by the form of its: (1) base, (2) shaft, (3) capital, (4) architrave.
6. The simplest molding used by the Greeks is the: (1) ovolo, (2) cyma (3) recta, (4) astragal.
7. The profile of the echinus in early Doric capitals shows: (1) a straight line, (2) a flaring curve, (3) a gentle curve.
8. The ceiling of the cella was usually of: (1) stone, (2) wood, (3) terra-cotta, and the beams were supported by: (1) columns, (2) pilasters, (3) buttresses.
9. In the Doric order sculpture was used to decorate the: (1) triglyphs, (2) architecture, (3) metopes.
10. The sides of the abacus of the Corinthian capital are: (1) concave, (2) convex, (3) straight.

HISTORY

The True-False form of test has been used in several history courses in Columbia College, notably in History

9, 16, and 165. No exact studies have been made of the results, but these have been so obviously valid and helpful to the members of the department that the teaching staff has committed itself to the policy of using the new forms of questions as a regular part of the examination machinery in history courses. The consensus of opinion among the instructors, according to Professors B. B. Kendrick and H. J. Carman, is that the New Type examinations given during the year yielded accurate and comprehensive information on the achievements of students in parts of the assigned work which could not have been gained from the Old Type questions, and that a great deal of time was saved by the use of objective and mechanically scored tests. One of the examinations in History 9 adequately illustrates the kind of New Type examinations used in this department during the past year:

MID-YEAR EXAMINATION, 1922[1]

History 9

PART I—ONE HOUR

NOTE: If you are reasonably certain that a statement is correct, mark it +. If a statement, or ANY PART of it, is untrue, mark it —. If you are uncertain, make no mark at all. It does not pay to guess in this sort of examination. However, in order to obtain a good grade, you should answer at least 60 questions.

.. 1. Franklin, Jefferson, and Samuel Adams were the American representatives who negotiated the treaty of peace at the close of the Revolution.
.. 2. Maryland was the last state to ratify the Articles of Confederation.
.. 3. North Carolina and Rhode Island were not represented at the Federal Convention.
.. 4. The members of the Federal Convention were not authorized to draft a new constitution, but only to frame amendments to the Articles of Confederation.
.. 5. Nationally-minded delegates, like Madison, Wilson, and Patterson championed the interests of the large states.

[1] The complete examination from which these sample questions are taken comprised 80 questions.

.. 6. In the struggle for ratification of the Constitution, manufacturers and merchants in general favored it, while the farmers as a class were opposed to it.
.. 7. John Jay, while Chief Justice of the Supreme Court, went to England and negotiated the Jay Treaty.
.. 8. While Chief Justice, John Marshall, together with Pinckney and Gerry, was sent to France by President Adams to negotiate a treaty with the Directory.
.. 9. Influenced by Gallatin's Report on Manufacturers, President Jefferson, in 1807, advocated an embargo to encourage our infant industries.
..10. The Treaty of Ghent was a remarkable diplomatic triumph for the United States; Great Britain yielded on every important point.
..11. One reason why Webster so earnestly advocated the protective tariff in 1816 was that the bulk of the capital in his state was by that time invested in manufacturing.
..12. John Marshall in his Supreme Court decisions translated into the Constitution the principles of Jeffersonian Democracy.
..13. The "Bonus Bill" of 1817, providing that certain moneys accruing to the government from the National Bank be used for internal improvements, was opposed in Congress chiefly by representatives from the New England states.
..14. Despite the fact that the Missouri Compromise was favorable to the North, it was opposed by the great majority of the representatives from that section.
..15. Since Jackson received a majority of the popular vote in 1824, he was morally entitled to the presidency.
..16. The Dartmouth College case provided that the United States could not grant a charter.
..17. The original purpose of the Foote Resolution seems to have been to strengthen the existing political alliance between East and West.
..18. In advocating the American system, Clay sought to convince the farmers of the West and the manufacturers of the East that their interests were essentially identical, perhaps hoping in this way to arrive at the White House *via* an East-West coalition.
..19. One objection on the part of the Westerners to the National Bank was that by inflating the currency it tended to increase their debts as measured in terms of their products.
..20. Both Webster and J. Q. Adams were opposed in the main to the extension of the suffrage to the propertyless classes.
..21. Jackson's first reorganization of his cabinet was for the purpose of getting rid of the Calhoun faction.
..22. The political sun of Jackson went down brightly, and his chosen successor declared openly that he would gladly follow in the footsteps of his predecessor.
..23. The Independent Treasury system proposed to avoid concentration and monopoly such as the national bank had maintained.

..24. John Quincy Adams persecuted politically in 1828 played the rôle of political persecutor in 1840.
..25. Clay declared that President Tyler dared not resist him in his program for a national bank. In this notion he was mistaken.
..26. The South, with the exception of South Carolina, was solidly in favor of the annexation of Texas.
..27. The Anti-Masonic party set the example for national conventions for the nomination of presidential candidates.
..28. Although Jackson resolutely opposed J. Q. Adams, he nevertheless adopted the latter's Indian policy.
..29. "Fifty-four-forty or fight" was the exultant slogan of Webster in the campaign of 1832.
..30. By 1825 the old social and political differences between the lowland and upland sections of the Southern States had disappeared.
..31. The Whig party suffered defeat in 1844 largely because of the blunders of its principal leader, John Tyler.
..32. The financial interests of the East demanded a renewal of the independent treasury system in order to provide a sound paper currency.
..33. Taylor, a southern slaveholder, was nominated by the anti-slavery faction in the northern wing of the Whig party to get control of the party out of the hands of the Clay faction.
..34. The application of California for admission brought on a crisis which nearly caused the secession of certain southern states.
..35. The Compromise measures passed in 1850 consisted of the following provisions:

Admission of California as a free state.
Organization of Utah and New Mexico on principle of "squatter sovereignty."
Prohibition of interstate slave trade.
More stringent Fugitive slave act.
Readjustment of Texas boundary.

..36. The treasury was overflowing in 1853 and the debt was being rapidly reduced.
..37. Western politicians were anxious to get land subsidies for a railroad to extend to the Pacific and consequently favored the organization of new territories.
..38. The Ostend manifesto was issued as a declaration of our disapproval of England's movements to suppress the slave trade.
..39. Charles Sumner was assaulted because of remarks which he made in defense of John Brown's raid.
..40.[1] Lincoln, a decided abolitionist, held the doctrine that the union could not endure half slave and half free.

[1] To economize space only 40 of the 80 questions of the History 9 test are reproduced here.

ENGLISH

Between courses in English there seems to be no common denominator beyond some kind of concern with books. Nor is the end to be achieved by this concern invariably defined. A course in literature may, in effect, be a course in philosophy, or in history, or in philology, or ethics or æsthetics—or in "appreciation": a course in any or all of these. As a result of this lack of uniformity of end and variety of emphases, the grades given in literature courses seem to be more capricious and difficult to interpret than any other courses given in college.

In courses that require any bulk of reading, and especially in large classes where, because of numbers, the instructor finds it impossible to follow up, by personal contact, the work of each student, the conventional examinations are obviously inadequate to test fully and specifically what reading each student has done. The student body is competently alive to this inadequacy, and not infrequently gambles on converting little work into high grades.

The True-False type of examination has shown itself to be an effective instrument for checking up on the scope and attentiveness of a student's reading. Mr. Raymond M. Weaver, Assistant Professor of English in Columbia College, has adapted this type of examination under circumstances and with results that are typical. His experiment is an admirable example of the intelligent and successful use of the New Type questions in a subject matter for which, on *a priori* grounds, it seemed to some critics that the objective type of questions were wholly unsuitable. In a memorandum kindly furnished to the writer, Professor Weaver thus outlines his experiment and the conclusions derived from it:

"At the beginning of this year I found myself facing what looked like an impossible state of affairs. In my English 15-16 course were 150 students. My reading list was long. Of each student I wanted to know what reading in the course he had done, and in what manner and with what vividness this reading had impressed him. The class was far too large for me to attempt to learn these facts through conferences. Frequent quizzes of the essay type were unmanageable. The hours necessary to examine the papers would have been enormous in number. Infrequent quizzes of any type I knew had proven themselves merely futile.

"With a little technical assistance, I made out a True-False examination based on the content of the lectures and the required reading. This experiment was completely salutary. Such a quiz, of a hundred propositions, given monthly, affords me a reliable index of how much each student has gained of the informational content of the course. Nor has any student, to my knowledge, questioned the justness of these ratings. These quizzes have, moreover, stiffened the effort of the class: it is no longer possible for a student to float to an 'A' grade on 'gush' and hollow generalizations.

"And these happy results were gained with a disproportionately small cost of time and effort to me. The making of such a quiz takes about four hours—very much less time than would be required to examine a quiz of the essay type. Through the coöperation of Dean Hawkes, these objective quizzes are scored by clerks in his office: a very desirable arrangement.

"The final grade is not, however, based solely on these quizzes. For a student might be perfect in his knowledge of the content of a book, and at the same time be dead to its essential meaning. As the best-known makeshift

for sampling the color of a student's mind, a term paper is required, and a final examination of the essay type is given. With this variety of evidence from each student, I feel—and this for the first time—that the grades in English 15-16 have some significance and reliability."

Sample questions from the New Type quizzes used by Professor Weaver are reproduced below:

Quiz English 15

DIRECTIONS: A few of these questions call for dates. Write the dates on the dotted lines at the left. The other questions are statements which you are to mark as true or false. If a statement is true, put a plus sign (+) on the dotted line at the left. If a statement is false, put a zero (0) on the dotted line at the left. A statement is considered false if any part of it is false.

.. 1. Æschylus was born in the year?
.. 2. Euripides and Sophocles both died in the year?
.. 3. Cicero wrote during the Silver Age.
.. 4. The Minoan Age was after the Dorian Invasions.
.. 5. Aristotle died in the year?
.. 6. Socrates died in the year?
.. 7. Sappho was contemporary with Nebuchadnezzar.
.. 8. Solon was contemporary with Thespis.
.. 9. Sappho was contemporary with Jeremiah.
..10. Sophocles' surviving plays were all written before he was forty.
..11. Sophocles fought in the battle of Marathon.
..12. Pericles fought many successful battles against the Persians.
..13. The Romans borrowed from the Greeks the use of rhyme.
..14. Julian the Apostate was the first Christian Emperor.
..15. In the Third Century B.C., Alexandria was noted for its scientists.
..16. Theophrastus wrote a book of pastorals.
..17. Theocritus was the father of botany.
..18. Bion and Moschus were Alexandrian scientists.
..19. Greek comedy, according to Aristotle, grew out of the dithyramb.
..20. Plato and Aristotle had fewer followers among the Romans than had Zeno and Epicurus.
..21. Helen and Klytaemestra are sisters.
..22. Io and Ion were both presented in Greek tragedy as suffering because of the apparent injustice of the gods.
..23. In the *Hippolytus* Aphrodite destroys the children of Artemis.
..24. Rather than be faithless to Medea, Jason fled from Corinth.

..25. Achilles bore the death of Patroclus with no outward show of emotion.
..26. Helen's appearance is closely described by Homer.
..27. On a nocturnal spying expedition the two Ajaxes killed Rhesus and his followers.
..28. Eurycleia recognized Odysseus by means of a mole on his neck.
..29. Æolus had no children.
..30. In the *Symposium*, Socrates compares sophists with cooks.
..31. Diomedes was a prophetess who taught Socrates the nature of love.
..32. In the *Symposium*, Agathon came drunk to the banquet.
..33. Aristophanes wrote a comedy in ridicule of Socrates.
..34. Plato represents Socrates and Aristophanes as being friends.
..35. Euripides has been frequently praised for the eloquent silences of his characters.
..36. Phædra was loved by Hippolytus.
..37. The death of Antigone results in the suicide of Creon's wife and son.
..38. The *Iliad* tells the story of the Trojan horse.
..39. The plots of very few of the surviving Greek tragedies are to be found in Homer.
..40. Medea killed her children.
..41. In Æschylus, Klytaemestra denies her guilt when accused of murdering her husband.
..42. In the *Eumenides* (Furies) Klytaemestra's ghost appears broken with remorse.
..43. Calypso was a mad prophetess.
..44. We have no Sophoclean tragedy in which, as in the *Prometheus* and *Eumenides*, the characters are immortals.
..45. Ismene is a cruel self-willed girl.
..46. Thucydides wrote a history of the wars with the Persians.
..47. In the *Memorabilia*, Xenophon comments at length on Plato's discipleship for Socrates.
..48. Æschylus wrote the *Acharnians*.
..49. Aristophanes wrote the *Persians*.
..50. Hecuba was Queen of Lesbos.
..51. Athens did not reach the height of her glory until after the Peloponnesian war.
..52. Roman literature is noted for its originality.
..53. Hermione was Helen's daughter.
..54. The details of the death of Ulysses as told by Dante are almost entirely as found in Homer.
..55. The Greek word *arete* is adequately translated into English as *virtue*.
..56. Virgil was not thought much of as a poet by his contemporaries.
..57. Œdipus died a miraculous death.
..58. According to Plato's acount, Alcibiades was present at Socrates' death.
..59. Socrates is famous for the dialogues he wrote.

..60. Pindar was a Roman poet.
..61. Terence wrote comedies.
..62. The *Furies* violates the unities of time and place.
..63. Herodotus was a contemporary of Sophocles.
..64. Calliope was Hecuba's daughter.
..65. Solon was an enthusiastic admirer of Greek tragedy.
..66. Plotinus was an Epicurean.
..67. Epicurus was a materialist.
..68. In the *Phaedrus*, Socrates tells how he frequently walks in the country to restore his faith in men.
..69. Æschylus' plays were presented in the theater of Dionysius.
..70. The audience that witnessed the plays of Sophocles sat on wooden bleachers.
..71. Aristophanes attributes great personal beauty to all the disciples of Socrates.
..72. Xanthius and Hercules were stock characters in Greek comedy.
..73. Strepsiades is a character in the *Clouds*.
..74. In Sophocles, Electra is outwardly married to a peasant.
..75. Hephæstus (Vulcan) sympathizes with the sufferings of Prometheus.
..76. Phidippides had a passion for horse-racing.
..77. Attis was loved by Sappho.
..78. Astyanax was the son of Achilles.
..79. In Euripides, Electra's hatred for Ægisthis was softened by her realization of his nobility of character.
..80. Praxiteles was the father of medicine.
..81. Galen advised Helen to return to Sparta and so stop the war.
..82. Andromache killed her children.
..83. Achilles showed no mercy to the father of dead Hector.
..84. Euripides was noted for his orthodoxy.
..85. Plautus borrowed his plots from the Greeks.
..86. Lucretius believed in rewards and punishments after death.
..87. For all practical purposes, the knowledge of Greek in the Middle Ages died with Boëthius in 525 A.D.
..88. Rome was the birthplace of Neo-Platonism.
..89. Quintilian was a Roman rhetorician.
..90. In *Sophocles*, Œdipus protests to the end that he suffers through no guilt of his own.
..91. Tiresias was a seer and prophet.
..92. Terence's comedies were not acted until the time of the Renaissance.
..93. Virtue, according to Aristotle, consisted in following the golden mean.
..94. Socrates taught that there is a close positive relationship between knowledge and virtue.
..95. Plato flourished after the great period of Athenian supremacy.
..96. Tragedy, according to Aristotle, purges the soul by pity and terror.
..97. In the Athenian theater, no character was ever acted by a woman.

..98. In the Athenian theater, the action was presented in the orchestra, not on a raised stage.
..99. Aristophanes repeatedly praises "the good old days."

CIVIL ENGINEERING

The New Examination form has been tried out in Civil Engineering, but no study of the results has been made. The examination is reproduced here, in part, as a further illustration of the wide applicability of the objective type of question. In the opinion of Professor Finch, who constructed and used the examination, these new types of questions give the same satisfactory results that they give in the other subject matters mentioned above.

MIDYEAR EXAMINATION, JANUARY, 1922

Hydraulics: C. E. 175

Part I

Mark such of the following statements as you believe to be correct + and those which are wrong —.

(Correct answers count two points, incorrect answers take away two points, and unanswered statements take away one point.)

.. 1. The coefficient of contraction for an orifice is always less than the coefficient of discharge.
.. 2. Stream line flow conditions occur only for relatively low velocities.
.. 3. Velocity of approach is a relatively unimportant factor in the majority of hydraulic measurements
.. 4. The daily records of stream flow published by the Government are computed directly from current meter measurements.
.. 5. The Chezy formula is best suited to problems involving the flow in ditches, canals, and non-pressure conduits of large size.
.. 6. The true kinetic energy of a stream is not always obtained by using the mean velocity.
.. 7. The equation $P = wAhcg$ is true only for plane surfaces.
.. 8. The loss of head due to a sudden increase in the size of a pipe is generally greater than that for a corresponding decrease.
.. 9. Submerged floats are generally used for flow measurement in deep slow-flowing streams and for current investigations.

..10. The best type of pipe to use for large size, high pressure lines is cast iron.
..11. The Pitot tube gives more satisfactory results in the measurement of high velocities.
..12. The maximum velocity of flow in a stream is usually found at a depth of six-tenths the depth of the stream.
..13. Wood will last for a long time if kept constantly saturated with water, hence wood stave pipe is extremely durable.
..14. A nozzle on the end of a pipe line acts to reduce the discharge and the pipe velocity but increases the terminal velocity.
..15. A weir is the most accurate device available for measuring the flow of a stream.
..16. The mean velocity of flow in a pipe will be slightly less at points where bends occur than at other points in the line.
..17.[1] Hemi-spherical bottoms are used for low tanks to increase their storage capacity.

PART II

Answer by simple and brief statements the following:

1. The loss of head in friction in a pipe depends on the following experimental laws: .....
.....
2. On what law is Bernoulli's Theorem based? .....
.....
3. The current meter depends on what assumption? .....
.....
4. The advantages of continuous wood stave pipe are: .....
.....
5. The equation for discharge through a Venturi meter is based on what law or laws? .....
.....
6. What methods are available for measuring the discharge of a nozzle? .....
.....
7. Name in order of accuracy three methods of measuring the flow of a stream: .....
.....

PART III

Cross out statements which are not true.

1. The pressure at any point due to gravity in a liquid at rest depends on the viscosity of the liquid, the depth, the character of the surface of the containing vessel, the density of the liquid, the volume of liquid.

[1] To economize space only 17 of the 33 questions of this test are reproduced here.

2. The vertical distance from the water surface to the center of pressure of a submerged, inclined plane surface depends on the unit weight of the liquid, the angle of inclination of the surface, the area of the surface, the location of the center of gravity of the surface, the weight of the displaced liquid, the shape of the surface.
3. The Cippoletti weir has the following special advantages: accuracy, constant coefficient, simplicity of computation, greater discharge than suppressed weir of same size.
4. Concrete pipes have the following advantages: low resistance to flow, great compressive strength, durability, will allow less leakage than other kinds, suited to high external pressure.
5. The factors entering the approximate formula for the pressure due to water hammer allow for the rate or time of closing the valve, the thickness of the pipe, the velocity of pressure transmission, the velocity of flow in the pipe, the diameter of the pipe, the slope of the hydraulic gradient.

Part IV

Use of slide rule and two blue-print diagrams permitted. Return question paper with your answer book.

1. Give a brief description such as would be suitable for an Introduction to the subject of Hydraulics, noting what it covers, its main divisions, and their general characteristics such as subject matter and general method of attack, etc.
2. The depth of water behind a masonry dam is sixty feet. Its back face is vertical for twenty feet, but the lower forty feet has a slope upstream of ten feet. Find the overturning moment of the water pressure about the back toe.
3. Using the average coefficients for orifices find the discharge and the velocity both at the contracted section and in the plane of the orifice for a circular orifice 4 inches in diameter under a ten-foot head.
4. A circular sewer flows one-half full. It is four feet diameter concrete pipe with a slope of one foot per one thousand. What is the probable discharge?
5. Make a drawing, with vertical scale very much exaggerated, showing the hydraulic gradient for a pipe line which leaves a reservoir at a depth of 12 feet below the water surface, runs 2,000 feet with a diameter of 24 inches and is then reduced to 12 inches for 1,000 feet, discharging under a head of 30 feet. No exact figures required.
6. Assuming $C = 100$ find the discharge of the above line.

CHAPTER THIRTEEN

Summary

In the preceding pages we have reviewed certain facts concerning two classes of tests, the intelligence test and what we have called the New Type Content Examination. On the basis of experimental evidence we have learned something of the values and applications of these two classes of measuring devices. We have learned that the intelligence test is the peerless admission criterion when used in conjunction with other available criteria. In the matter of predicting academic achievement, the intelligence test stands without a serious rival. The only other available admission criterion which approaches the intelligence test in power of prediction is the New York State Regents Examination, which is as much an intelligence test as it is a content examination, and which gives comparable results for all students who take it. By the use of the intelligence test many candidates who have made satisfactory or passing grades in high school, but who have not the superior intellect requisite to success in college, are saved from discouraging failure and waste of time and money by being advised to undertake work more suited to their talents. We learned that the intelligence test gives a solid foundation for building up a system of differential admission, whereby students would not be admitted to branches of the university for which their intellectual powers render them clearly unfit.

The use of the intelligence test makes it possible to select the very superior minds among entering students for early attention. In Columbia College all entering students who make intelligence scores above one hundred have a personal interview with the Dean, who assigns them to

the special attention of one or more of the instructors, with the understanding that everything is to be done that can be done to stimulate and guide the ambitions and efforts of such men to become acquainted with and increase the store of human learning and art. Very superior minds are usually (but not always!) discovered sooner or later; the advantage of the intelligence test is that they are sifted out at once for the special attention which they deserve and which the higher interests of society demand that they should get.

While the very superior minds rightly receive special attention, their nurture does not rob the less superior of any opportunity to develop. One of the greatest advantages of the intelligence test is that it enables the administration to adjust academic load to individual capacity so much more precisely than heretofore that many mediocre students, who would otherwise have dropped out of college, are saved from failure during the first year, and thus are allowed to go on through college with as much speed and profit as their intelligence allows. It is hardly necessary to mention the other administrative uses of the intelligence test referred to in this book, such as in the classification of students for instructional purposes, in the diagnosis of special cases in the Dean's Office, as pertinent information in all disciplinary cases, in advising students on broad vocational lines, and in research.

Nor is it necessary to recall here in detail the facts concerning the New Type Content Examination which form the subject matter of the last half of the book. The inadequacy of the traditional examining methods and the efficiency and economy of the New Method have been placed beyond the realm of doubt. Experimental evidence has shown that the New Type tests measure what the instructors desire to measure more accurately than the

Old Type examinations, and at less cost. The objectivity and comprehensiveness of the New Type tests give them very important administrative and pedagogical values not possessed by the traditional examination methods. Since the New Type tests measure specialized achievement in very large classes and groups of classes on an objectively comparable basis, the door is opened for genuinely scientific research in the field of collegiate instruction and pedagogy. And since they measure defined achievement in special fields of intellectual endeavor thus objectively and reliably, the basis is laid for concrete educational and vocational advice.

The articulation of exact measurements with the general administration of the college may be best indicated by outlining the scheme now in almost complete operation in Columbia College. It is, of course, unnecessary to remind the reader that the experience of a particular college is described here only for the sake of presenting a concrete example of the application of principles which are equally valid for any college, normal school, or high school of superior rank, whatever its geographic location may be.

In addition to the usual procedure of getting into college, such as making formal application for admission, furnishing certified secondary school records, health records, character recommendations, life history, etc., all students entering Columbia College for the first time will be required to submit to two examinations, (1) the intelligence examination, and (2) the content placement examination. The scores on both these examinations in analyzed form will be put on the permanent record cards of the students. The intelligence test has been described in Chapter Three and illustrated in Appendix I. The content placement examination, in its final form, will consist

of a battery of New Type objective tests for achievement in English, Mathematics, Modern Languages, Science, and History, organized in accordance with the following abbreviated specifications:

I. ENGLISH
   1. Vocabulary
   2. Spelling
   3. Grammar and Rhetoric
   4. Literary Information (including scansion)
   5. Composition (narrative and expository)
   6. Reading

II. MATHEMATICS
   1. Algebra (Elementary, Intermediate, Advanced)
   2. Plane Geometry
   3. Solid Geometry
   4. Trigonometry

III. MODERN LANGUAGES
   1. French
   2. Spanish
   3. German
   4. Italian

IV. SCIENCES
   1. Physics
   2. Chemistry
   3. Biology
   4. Botany
   5. Geology

V. HISTORY
   1. Ancient
   2. Mediæval and Modern
   3. English
   4. American

Every candidate will be required to take at least the following sections of this battery:

I. English, complete;
II. Mathematics, Algebra and Plane Geometry;
III. One Modern Language

Additional sections may be required of or chosen by the candidate, depending on a variety of considerations and conditions. From four to seven hours are required to give the battery to a group of five to eight hundred candidates. Each section of this content placement battery is validated and standardized in terms of the existing standards in Columbia College.

As a result of the use of these two examinations, the administration of the college knows before the student enters college two things about the student which heretofore have been learned very inexactly, if at all, several months after the student has been in college, viz.: (1) the student's intellectual capacity; and (2) the student's analyzed achievement in the fundamental undergraduate subject matters. Moreover, this knowledge is in terms which are not only comparable as between all students in the college, but in terms which have been given meaning for the administration by direct and objective reference to standards maintained in the college. Thus we know at once, (1) the ultimate limit of achievement which the candidate is mentally capable of reaching in Columbia College; (2) the speed which he is mentally capable of maintaining; and (3) just the points at which he can most advantageously begin his studies in the fundamental subject matters as taught in Columbia College. On the basis of this information, both vocational and educational advice becomes more precise. A relatively low intelligence score, coupled with a poor showing on those sections of the content placement examination devoted to mathematics and science, would certainly make it inadvisable to encourage a boy to enter the School of Engineering. Experience has shown that boys below a certain level in intelligence who have difficulties with elementary mathematics and science, rarely if ever succeed in getting

through the Columbia School of Engineering. The same is true of the Schools of Medicine and Law. Nor would a boy with a relatively low intelligence score be allowed to register for more than fourteen or sixteen hours of work in one semester. By contrast, boys with very high scores are not only allowed but encouraged to attempt heavy programs; those making scores above one hundred normally pass the equivalent of from nineteen to twenty-two points each semester in terms of real academic achievement. Those who are found deficient in any respect, for example, in English grammar, composition, or spelling, are advised to register in University Extension courses specially designed to overcome such defects, either before they are allowed to register for the regular college work in English, or contemporaneously with such regular work.

The permanent record card mentioned above is kept in the Dean's Office and is designed to present a complete up-to-date summary of all that is known about each student in the college. The flexibility of the card (see last section of Chapter Three, pp. 58-62) makes possible a permanent record of every item of information of whatever kind which comes to the Dean's Office. From the viewpoint of continuing vocational and educational advice the great advantage of this card, in addition to the bringing together in one view of all the pertinent information available, is that it always presents to the administration a complete transcript of the student's academic work in the college, analyzed by subject matter and expressed in terms of scores on the new objective content examinations. Whenever the records are based on any other type of examination, the fact is noted. The keeping of such a card is now possible in every college, normal school, and high school in the country at a cost which is negligible

in comparison with its manifold uses in the administration of an institution of higher learning and its immeasurable ultimate returns in terms of "student-accounting." Let us not forget that the thing which makes such a card valuable or worthless is the character of the records which it contains. The keeping of a card showing a student's "marks" has been possible for many years; but "marks" having a reliability coefficient of only 0.30 to 0.50 have hardly been worth preserving. The Columbia College card is valuable and useful because it contains intelligence scores with a reliability not lower than 0.85, and because it contains analyzed educational scores with reliability coefficients ranging from 0.80 to 0.95, and because both types of scores are expressed in units whose meaning will always be clear to the administration of the college.

The making of intelligence tests is and will remain in the hands of experts. Standardized intelligence tests are now on the market suitable for every grade of school from kindergarten to graduate college department. But a part of the work of making valid content examinations will probably always rest upon local teachers and administrators; hence it seems desirable to summarize the technique of constructing and interpreting the results of such tests. What is the general procedure in making and interpreting the results of a valid examination in, let us say, plane geometry?

In making out any examination, the first question which must be answered is, What is it that we wish to measure? He who would make an examination in geometry is no exception to this rule. What is geometry? The answer given by the College Entrance Examination Board may be judged from the character of the examinations in geometry which the Board has been using for over twenty years:

COLLEGE ENTRANCE EXAMINATION, 1921

PLANE GEOMETRY COMPLETE

1. Prove: If two parallel lines are cut by a third straight line, the alternate interior angles are equal.
2. Prove: An angle inscribed in a circumference is measured by half the intercepted arc.
3. Prove: The areas of two similar triangles are to each other as the squares of two corresponding altitudes.
4. In an isosceles triangle, ABC, the sides AC and BC are each twenty-five inches long, and AB is thirty inches. How far is A from the side BC?
5. A point P lying within a quadrilateral is connected with each of the vertices. Show that the sum of the length of these four lines is least when P lies at the intersection of the diagonals.
6. Two circles, C and C', are tangent to each other externally, and each is tangent to a straight line, L. The straight line through the centers of C and C' makes an angle of 30° with L. If the radius of the larger circle is three inches, find the radius of the smaller circle.

The definition which this examination implies seems, on careful consideration of the goals set up by the mathematical associations, very inadequate. A partial inventory of the more or less distinct groups of experiences which students are supposed to have in a course in plane geometry would include the following:

(1) General information about geometry, its nature, its terms, its materials, its divisions, its history, its articulation with other branches of mathematics, and with other sciences and arts:—what we might call the fundamental cultural experience, which after all is perhaps the most enduring benefit which the majority of students get out of the course.

(2) Definitions; that is, a skill in expressing ideas clearly and succinctly; the ability to understand and express exactly the meaning of terms, phrases, assertions, relations, postulates, axioms, etc.

(3) Converse relations.
(4) Quick geometrical reasoning without rigid proofs.
(5) Demonstrations of theorems and of original propositions.
(6) Applied problems and computations.
(7) Constructions with and without proofs.

In the College Entrance Board examination reproduced in full above there is not a single reference to the general nature, materials, divisions, or history of geometry; not a single definition is called for; the important logical training gained from the consideration of converse relationships is left untouched; constructions and computations are largely incidental, purely formal, and so easy as to afford no adequate measure of these important abilities.

The definition once agreed upon, the next step is to construct an examination comprehensive enough to satisfy that definition, objective enough to insure that competent examiners everywhere will secure comparable results, and reliable enough to make the results worthy of being used by administrators, instructors, and research scholars. The examination may be made comprehensive by the use of the New Type questions, whose brief recording processes make it possible to include a comparatively large number of each of several different kinds of questions. The brevity of the answers to the New Type questions not only rescues a very large part of the examination period from irrelevant activities and considerations on the part of the examinees, but at the same time renders the examination objective and reliable. To illustrate concretely the implications of the last two paragraphs, the following questions are suggested here merely as samples of forms of questions which might go into a valid examination in plane geometry:

Section I

GENERAL INFORMATION

Directions for Section I: Fill in the blank spaces in the following sentences so as to make sensible and significant statements. Put one word in each one and one-half inches of unbroken space. If, after thinking over a given sentence for a few moments, you do not see the point, omit it and go on to the next, going back to it later if you have time. Do all the easy ones first. Don't make sheer guesses, as you will be penalized for foolish or absurd completions.

1. Algebra may be called a ..... arithmetic.
2. Geometry is a branch of mathematics not primarily devoted to ..... as in arithmetic, nor to equations as in ....., but is principally concerned with ....., such as ....., ..., and circles.
3. That which has ..... and ..... without ..... is called, in geometry, a surface.
4. In geometry that which has length without ..... or ..... is called a .....
5. Solids, surfaces, and lines are called .....
6. A ..... has only ....., without length, breadth, or thickness.
7. The intersection of two lines is a ..... and the ..... of two planes is a .....
8. In geometry we think of a line as ..... by a moving point, a surface as ..... by a moving ..... and a ..... as ..... by a moving surface.
9. A line made up of two or more different straight lines is called a ....
10. Two angles that have the same vertex and a common side between them are called ....
11. A ..... curve lying in a plane, and such that all of its ..... are ..... distant from a fixed point in the ....., is called a circle.
12. Any portion of a circle is ....
13. Two complementary angles whose difference is 30° are .....° and .....°.
14. A line that cuts two or more lines is called a ..... of those lines.

Section II

GENERAL INFORMATION

Directions for Section II: Read these statements and mark each one at the left of its number with a plus-sign (+) if you think it is

true, with a zero (0) if you think it is false. Your score will be based upon these signs only; don't waste time writing anything else. First go through the list quickly and mark all that you know for certain at once; then go back and study out the harder ones.

Don't guess! A wrong response counts heavily against you. Omit those about which you know nothing.

A statement is true only if every part of it is true; a statement is false if any part of it is untrue.

.. 1. A piece of piano wire is a geometric line.
.. 2. The end of a very thin wire is a good example of a geometric point.
.. 3. Part of a straight line is called a segment of that line.
.. 4. Geometry treats of geometric figures of all kinds.
.. 5. Solid geometry treats only of figures that do not lie wholly in the same plane.
.. 6. Two lines which intersect are not in the same plane.
.. 7. A circle is a curvilinear figure.
.. 8. The opening between two straight lines drawn from a point is called a straight angle.
.. 9. A line drawn through a circle, terminated at each end by the circle, is a diameter.
..10. When a right or obtuse angle is bisected the two resulting angles are acute.
..11. If we bisect a reflex angle, we always form two obtuse angles.
..12. If two adjacent angles are equal, their sum is equal to two right angles.
..13. Vertical angles are always equal and adjacent angles.
..14. All the interior angles of an isosceles triangle are acute.
..15. The angles of an equiangular triangle are equal to the angles of every other equiangular triangle.
..16. No polygon can be cut in more than two places by one straight line.

Section III

Geometric reasoning without proof

Directions for Section III: (Exactly the same as for Section II).

.. 1. Two triangles are equal if the three sides of one are equal to three sides of the other.
.. 2. Two right triangles are congruent if any two sides of one are equal to two sides of the other.
.. 3. A straight line parallel to two straight lines is in the plane of those two lines.
.. 4. If two parallel lines are cut by a transversal the two interior angles on the same side of the transversal are complementary.
.. 5. When two lines are cut by a transversal, if the exterior-interior angles are supplementary the lines are parallel.
.. 6. An exterior angle of a triangle is larger than either of the opposite interior angles.

.. 7. **A diagonal is a straight line joining any two vertices of any figure.**
.. 8. Two angles whose sides are parallel each to each are equal.
.. 9. If the opposite sides of a quadrilateral are equal, the figure is a rectangle.
..10. The diagonals of a quadrilateral bisect each other.
..11. The line which joins the midpoints of two sides of a triangle is parallel to the third side, and is equal to one-half the third side.
..12. No concave polygon is equiangular.
..13. Each angle of a regular polygon of n sides is equal to $2(n-2)$ right angles/n.
..14. The sum of the exterior angles of a polygon, made by producing each of its sides in succession is equal to four straight angles.
..15. The proof of a locus problem is complete when it has been shown that any point in the supposed locus satisfies the conditions laid down in the problem.
..16. In an isosceles triangle the medians drawn to the equal sides are equal.
..17. The bisectors of the equal angles of an isosceles triangle form, together with the base, an isosceles triangle.
..18. The bisectors of two consecutive angles of a parallelogram are perpendicular to each other.
..19. If two straight lines intersect at right angles, any point in either is equidistant from the extremities of the other.
..20. If either diagonal of a parallelogram bisects one of the angles, the sides of the parallelogram are all equal.
..21. If two sides of a quadrilateral are parallel, and the other two sides are equal but not parallel, the sums of the two pairs of opposite interior angles are equal.
..22. The sum of the distances of any point from the three vertices of a triangle is greater than one-half the sum of the sides.
..23. If the opposite sides of a hexagon are equal and parallel the diagonals that join opposite vertices meet in a point.
..24. The lines joining the midpoints of the consecutive sides of any quadrilateral form a square.
..25. If the diagonals of a parallelogram are equal all the angles of the parallelogram are equal.
..26. If the diagonals of a parallelogram are equal the parallelogram is a square.
..27. A line through the center that bisects any chord bisects the arc subtended by that chord.
..28. All the lines bisecting a chord coincide.
..29. If two chords intersect and make equal angles with the diameter through their point of intersection, these chords are equal.
..30. All equal chords of a circle are tangent to a concentric circle.
..31. If two circles intersect, the line of centers is parallel to their common chord.

..32. If, while approaching their respective limits, two variables are always equal, their limits are equal.
..33. A portion of a plane bounded by an arc of a circle and its chord is called a sector of the circle.
..34. In plane geometry we have limiting forms as well as limiting values.
..35. In the same circle or in equal circles two central angles have the same ratio as their intercepted arcs whether the arcs are commensurable or not.
..36. An angle whose vertex is on a circle, and whose sides are chords, is measured by half the intercepted arc.
..37. If an angle is inscribed in a semicircle so as to form an isosceles triangle with the diameter as a base, this angle is greater than any other angle inscribed in the same semicircle which does not make an isosceles triangle with the diameter as a base.
..38. Angles inscribed in the same segment or in equal segments are always equal.
..39. Every parallelogram inscribed in a circle is a square or a rhombus.
..40. Any trapezoid may be inscribed in a circle.
..41. An angle formed by a tangent and a chord drawn from the point of contact is measured by half the intercepted arc.
..42. To construct a triangle with sides two inches, three inches, and six inches is called a determinate problem.
..43. To construct a quadrilateral of given sides is an indeterminate problem.

Section IV

COMPUTATIONS

Directions for Section IV: Find the answers to these problems as quickly as you can and write them on the dotted lines. Do the easy ones first and go back to the harder ones later. Use the margins of this page to figure on if you need to.

1. The diameter AB is the base of a triangle formed by inscribing an angle ACB in a semicircle. Angle B equals 30°. Angle A equals .....
2. Angle A equals 34°. What is the magnitude of angle B its complement? .....
3. How many degrees in the conjugate of an angle equal in magnitude to the sum of the interior angles of any triangle? .....
4. How many degrees in the supplement of the angle formed by bisecting one of the base angles of an equilateral triangle? .....
5. How many degrees in an angle which equals its own complement? .....
6. How many degrees in an angle which equals one-third its own conjugate? .....

7. If angle X plus angle A = 4 angle A minus angle X, find the value of X. .....
8. If angle X plus Y equals 45° and X minus Y equals 35° what are the values of X and Y? ....
9. The base of an isosceles triangle is x; one of the sides equals 2x; what is the other side? .....
10. In a certain isosceles triangle angle A is twice angle B; what are the values of angle A? Angle B? Angle C? ....
11. How many degrees in each angle of an equiangular triangle? .....
12. One angle in a rhombus is 30°; what are the magnitudes of the other three angles in order? ....
13. In a certain rhomboid two consecutive sides are twelve inches and fifteen inches in length; how long is the side opposite the twelve-inch side? .....
14. In a certain parallelogram ABCD the diagonals intersect at O. The segment AO of the diagonal AC is twelve inches. What is the length of diagonal AC? .....
15. In a triangle ABC, AB equals twelve inches. How long is the line joining the midpoints of AC and BC? .....
16. What is the sum of the interior angles of an octagon? .....
17. How many sides has a regular polygon each angle of which is 140°? .....
18. How many sides has a regular polygon each angle of which is one and three-fourths right angles? .....
19. How many degrees in each exterior angle of a regular hexagon?
20. A piece of piano wire six feet long is stretched taut from an upper corner of a rectangular window to a point in the lower side. A small ant crawls up the wire three feet. How far is he from the vertex of the angle of the window opposite the wire? .....
21. The base AB of an isosceles triangle is twelve inches. From the midpoint of one of the equal sides a perpendicular is dropped cutting AB in O. How long are AO and OB? ....
22. In a certain parallelogram the diagonals are equal. Find the magnitudes of the four angles. .....
23. In a circle the radius of which equals four inches, the tangents at the extremities of a chord AB are parallel; how long is AB? .....

Section V

converses

Directions for Section V: Read these statements, and mark each one at the left of its number, with a plus-sign if you think

its converse true, with a zero if you think its converse is false. Then write out the converses of the first six in the blank spaces.[1]

.. 1. A square is a parallelogram.
.. 2. An equilateral triangle is equiangular.
.. 3. A square is a rectilinear figure having four equal sides.
.. 4. In an isosceles triangle the angles opposite the equal sides are equal.
.. 5. Two equal angles are two right angles.
.. 6. Two triangles are congruent if two angles and the included side of the one are equal respectively to two angles and the included side of the other.
.. 7. Two lines in the same plane perpendicular to the same line cannot meet however far they are produced.
.. 8. If two parallel lines are cut by a transversal the alternate-interior angles are equal.
.. 9. A parallelogram is a rectangle.
..10. A trapezium is a quadrilateral.
..11. In the same circle equal central angles intercept equal arcs; and of two unequal central angles the greater intercepts the greater arc.
..12. A diameter bisects the circle.

Section VI

DEFINITIONS

Directions for Section VI: Define the following terms as clearly, exactly, and succinctly as you can, remembering the beauty of brevity as well as the absolute necessity for completeness.

1. Median.
2. State the generalized proposition on congruency of triangles.
3. Synthetic method of proof.
4. Analytic method of proof.
5. Straight line.
6. Chord.
7. Common external tangent of two circles.
8. Measure of a quantity.
9. Ratio.
10. Commensurable magnitudes.
11. Inscribed angle.

Section VII

LOCUS PROBLEMS

Directions for Section VII: State in a legible sentence or phrase or indicate by a rough drawing, without proof, the following loci in a plane:

[1] After each of the first six elements space sufficient for the examinee's writing should be left.

1. The locus of a point two inches from a fixed point O.[1]
2. The locus of the center of a circle that rolls around another circle always just touching it.
3. Locus of the point of a round lead pencil as it rolls along a smooth desk top.
4. Locus of a point equidistant from the extremities of a given line.
5. Locus of a point equidistant from two given intersecting lines.
6. Locus of a point four inches from a given point and three inches from a given line.
7. Locus of a point one-half inch from one given point and three-fourths of an inch from another given point.
8. Locus of the midpoint of a chord of given length that can be drawn within a given circle.
9. Locus of the midpoint of a chord drawn through a given point within a given circle.
10. Locus of a point that is equidistant from two given intersecting lines and at a given distance from a given point.

Section VIII

CONSTRUCTIONS WITHOUT PROOF

DIRECTIONS FOR SECTION VIII: Without proof make the following constructions, indicating the method with lines and dotted lines.

1. Erect a perpendicular to a line AB at a given point P in the line.
2. Bisect a given line AB.
3. Erect a perpendicular to a line AB at B.
4. Bisect an angle.
5. From a given point in a given line draw a line making an angle equal to a given angle.

Section IX

CONSTRUCTIONS WITH PROOF

DIRECTIONS FOR SECTION IX: Construct and prove the following:

1. Construct a triangle when two sides and the included angle are given.
2. Construct a triangle when two sides and the angle opposite one of them is given.
3. Circumscribe a circle about a given triangle.
4. Through a given point P draw a tangent to a given circle.
5. Upon a given line as a chord, describe a segment of a circle in which a given angle may be inscribed.

[1] After each element at least four lines of space should be allowed.

Section X

PROPOSITIONS

Directions for Section X: Choose the easiest one first.

1. Prove: Two triangles are congruent if two sides and the included angle of the one are equal respectively to two sides and the included angle of the other.
2. Prove: Two right triangles are congruent if the hypothenuse and a side of the one are equal respectively to the hypothenuse and a side of the other.
3. Prove: In the same circle equal chords are equidistant from the center.
4. An angle formed by two chords intersecting within the circle is measured by half the sum of the intercepted arcs.

Section XI

APPLIED PROBLEMS

Directions for Section XI: Solve the easiest first.
(Any good problems applied to concrete situations will serve. It is, of course, understood that substituting numbers in a formula does not constitute a good problem.)

The reader must understand that these sample questions as put together here are meant as sample questions, and not as a sample of a complete examination. The divisions into sections are made merely to facilitate exposition; and the length of the sections as presented here has been determined entirely by the desire of the writer to illustrate the flexibility and wide applicability of each type of question.

The problem of preserving comparable standards in content examinations presents greater difficulties. It is comparatively easy, with the objective content examinations, to maintain comparable standards of achievement within one school, or among the schools of a city or state in which the curriculum is uniform or nearly uniform for all schools. But to make test results in Montana and Florida comparable to those in Maine and Arizona is more difficult. In some subject matters, whose content

is relatively fixed, or whose content varies within narrow limits from one school to another or from one state to another, a single standard examination (but having many equivalent forms) can be prepared by competent authorities which can be used everywhere, thus permitting the indication of specific achievement in terms of units which are everywhere the same and everywhere understood. Comprehensive examinations in some of these subject matters, such as Reading, Grammar, Spelling, Algebra, Geometry, French, Spanish, Ancient History, etc., either have been or are being prepared on the basis of careful syntheses of the syllabi for these courses used in representative schools in the United States. But for such courses as are given in the history, government, economics, and philosophy departments, each department and school will have to make out its own examinations and set its own standards. But the relations of these standards to standards in other institutions, or in the same institution at different times, can be learned indirectly through the relative achievements of the students in intelligence and educational tests which are standardized. Knowing, as we do know, the high relationship between intelligence and academic achievement, we can safely say that two large classes in elementary economics in two colleges having about the same average Intelligence Quotient will also reach about the same understanding of elementary economics in the same time and under the same general conditions of instruction. Therefore, if we find that one such college gives percentages of indicated grades as in line (1) and another as in line (2), we should be justified in assuming that, for all practical purposes, a grade of A in the first college is equivalent in terms of real achievement to a grade of B in the second, and that B in the first is equivalent to C in the second, etc.

	A	B	C	D	E	F
(1)	5	30	50	10	5	
(2)		5	30	50	10	5

Thus, comparability of grades in different schools can be secured for non-standardized New Type tests by comparing the distributions of such grades with the distributions of the grades which the same students have received on a standard and widely used intelligence test, or on any standardized educational test.

APPENDIX ONE

1. Sample Questions of the Thorndike Intelligence Examination for High School Graduates

Series of 1919, 1920, 1921, and 1922

It would be desirable to reproduce here, for illustrative purposes, an exact copy of the complete intelligence examination used in Columbia College, consisting of five booklets; but limitations of space allow only such minimal samplings as will be sufficient for the reader to understand the form, length, and difficulty of each different kind of test in the total examination. The first three of the five booklets are exactly the same in form, differing only in specific content; and Part III is exactly the same in form as Tests 1 and 2 in Part II. Therefore, it will be necessary to illustrate only one form of Part I and one form of Part II. The directions and usually the first, middle, and last questions of each test will be reproduced. The general directions for giving and scoring the total examination will be given in condensed form.

PART I

Form M

TEST 1

Directions: Look at these four words:

Forward Backward Sidewise Lengthwise

1. Spell the first word backward, writing the result here .....
3. Write the last word in the lower left-hand corner of this page.
5. Write the first word parallel to the left-hand edge of this page about one-quarter inch in from the right-hand edge of this page and about three inches down from the top of the page.

TEST 2

Directions: Look at each sentence. Think what it would be if the words were put in the right order. If it would be

true, draw a line under the word true. If it would be false, draw a line under the word false.

1. steep is walking up hard hill a.....true false
6. the not with the oil water mix will.....true false
10. Lincoln U.S. of the president first was the....true false

TEST 3

DIRECTIONS: Perform the operations indicated. Express all fractions in answers in lowest terms.

1. ADD $\frac{3}{4} + \frac{1}{8} + .25 =$

4 yr. 6 mo.
1 yr. 2 mo.
6 yr. 9 mo.
3 yr. 6 mo.
4 yr. 5 mo.

3. SUBTRACT $8.6 - 6.05007 =$ $\frac{7}{8} - \frac{2}{3} =$ $5\frac{7}{16} - 2\frac{3}{16} =$
8. DIVIDE $4\frac{1}{2} \div 7 =$

TEST 4

DIRECTIONS: Write the correct answers to these problems. Use the margins of the pages to figure on.

1. I bought 4½ yards of cloth, gave the clerk $2 and got 20 cents as correct change. What was the price of the cloth per yard?.....Answer.
6. How much faster does a train go that covers 20 miles in 15 minutes than a car which goes three-quarters of a mile in 45 seconds?.....Answer.
10. A company marched 120 miles in 5 days. How many times as fast must they march to cover 90 miles in three days?.....Answer.

TEST 5

DIRECTIONS: In each of the sentences below, you have a choice among four words. Draw a line under the one of these four words which makes the truest sentence.

1. Ivory is obtained from mines elephants reefs oysters.
6. Cheviot is a kind of silk cotton wool linen.
10. The color of oxygen is green brown yellow colorless.

TEST 6

DIRECTIONS: If the two words of a pair mean the same or nearly the same, write *s* after the dots. If they mean the opposite or nearly the opposite, write *o* after the dots as shown in the samples. If you cannot be sure, guess.

Samples good—bad.....o
little—small....s

1. Cold—Hot .....
6. Knave—Villain...
11. Agony—Bliss.....
16. Aggrandize—Belittle ..
20. Plenary—Complete....

TEST 7

DIRECTIONS: Make a cross before the best answer to each question.

3. If you break your leg while on the march, what is the thing to do?
   go lie down under the next tree
   fire off your gun and shout
   take a drink from your canteen
   ask some one to send an ambulance
5. Why should every one save some money? Because
   no one needs all the money he earns
   it is necessary to have money ahead for luxuries
   a time may come when he cannot work
   the Bible tells you to

TEST 8

DIRECTIONS: In the lines below, each number is gotten in a certain way from the numbers coming before it. Study out what this way is in each line, and then write in the space left for it the number that should come next. The first two lines are already filled in as they should be.

Samples (2, 4, 6, 8, 10, 12)
(11, 12, 14, 15, 17, 18)

1. 27, 24, 21, 18, ...
6. 2, 1, 5, 4, 8, 7, 11, ...
10. 9, 8½, 7, 6½, 6, 5, 4½, ...

TEST 9

DIRECTIONS: In each of the lines below, the first two words have a certain relation. Notice that relation and draw a line under the *one* word in the parenthesis which has that particular relation to the third word. Begin with No. 1 and mark as many sets as you can.

Samples sky—blue: grass—(grow, green, cut, dead)
fish—swims: man—(boy, woman, walks, girl)
day—night: white—(red, black, clear, pure)

1. pupil—young: teacher—(school, old, woman, child)
6. engine—wheels: sled—(slidė, snow, runners, iron)
14. moon—night: sun—(day, shine, warm, large)
20. foot—leg: hand—(fingers, head, arm, grasp)

TEST 10

DIRECTIONS: Draw a line under the largest number and also under the smallest number in every column, as shown in the sample.

Sample	(1)	(5)	(9)
34	45	502	2170
87	46	454	9000
79	77	135	7997
68	73	550	7752
25	48	845	3330
82	54	749	8631
27	70	235	6126
30	89	563	7428
21	42	865	3455
24	21	780	8191

TEST 11

DIRECTIONS: Write the letter *f* before each of these statements which could not possibly be true.

1. Our horse grew so tired that finally we were compelled to walk up all the hills.
3. In some states there is a law forbidding a man to marry his widow's sister.
5. Owing to the lack of ready money, the shrewd financier was very frequently unable to take advantage of the rare bargains offered.
8. We turned eastward with the ruddy glow of the setting sun shining in our eyes.

TEST 12

DIRECTIONS: Fill in with conclusions which can be drawn correctly from the given facts in each set, as shown in the first line.

Given Facts	*Conclusions*
K is older than B	therefore E is older than F
A is as old as E	therefore K is ..... A
F is younger than A	therefore K is ..... E
K is younger than F	therefore A is ..... B

TEST 13

DIRECTIONS: Look at No. 1 on this page long enough to get it in mind. Then turn over the leaf, find the same thing on page 10 or 11, and write 1 under it. Then look at No. 2 on this page, turn over the leaf, find it, and write 2 under it. Do the same with No. 3, No. 4, etc.

1. 2. 3. 4. 5.

92351647
11

80042000
12

75923198
13

75953198
14

60254126
15

(On the next two pages are four groups of twenty-five figures each. These are omitted here.)

PART II

Form D

TEST 1*a*

DIRECTIONS: Read this paragraph and then write answers to questions 1, 2, 3, 4, 5, and 6. Read it again if you need to. Each entirely correct answer counts 3; a partly correct answer counts 1. Each wrong answer counts 1 or 2 off from your score.

The career of Sir John Macdonald must be considered in connection with the political history of Canada and the conditions of its government during the latter half of the nineteenth century. Trained in a school where the principles of responsible government were still in an embryonic state, where the adroit management of coalitions and cabals was essential to the life of a political party, and where plots and counter-plots were looked upon as a regular part of the political game, he acquired a dexterity and skill in managing men that finally gave him an almost autocratic power among his political followers. But great personal qualities supplemented his political dexterity and sagacity. A strong will enabled him to overcome the passionate temper which marked his youth, and later in his career a habit of intemperance, which he at first shared with many public men of his time. He was a man of strong ambitions, but these were curbed by a shrewd foresight, which led him for a long time to submit to the nominal leadership of other and smaller men. Politics he made his business, and to this he devoted all his energies. He had the gift of living for the work in hand without feeling the distraction of other interests. He had a singular faculty for reading the minds and motives of men, and to this insight he perhaps owed the power of adaptability (called by his opponents shiftiness) which characterized his whole career. To this power the successful guidance of the Dominion through its critical formative period must be ascribed.

1. Which two of these statements seem to you to be most in accord with the account given by the paragraph?
   a. Macdonald was a political boss.
   b. Macdonald was naturally an autocrat.
   c. Macdonald was a great statesman.
   d. Macdonald was a man of very wide interests.
2. What quality led Macdonald to let less able men than himself think that they bossed him? .....
3. What two faults in Macdonald's character are admitted by the writer of the paragraph? .....
4. What two names are given in the paragraph to the same trait of character, one being derogatory? .....
5. What is the "Dominion" mentioned in the last line? .....
6. Which of these would you put as the most important cause of Macdonald's success? Underline it.
   a. A strong will.
   b. Knowledge of human nature and ability to manage men.
   c. Shrewd foresight.
   d. Strong ambitions.
   e. Willingness to make a business of politics.

(Test 1*b* is exactly like 1*a*, except with different content, and is therefore omitted here.)

TEST 2

DIRECTIONS: On each line of dots write the word that makes the best meaning. Each sentence completed with entire correctness counts 3. A substantially correct completion will count 1. 2 will be subtracted from your score for each foolish or irrelevant completion of a sentence.

1. As ..... the treasure he had come to seek, probably it existed ..... in his own .....
4. Sometimes the cartoon ..... in a single ..... the characteristics .... man or ..... nation.
10. Many ..... are unveiled, many more ..... by the fear of publicity.
12. It is only as ..... that suspense of ..... between Björnson and Ibsen is for a moment .....

TEST 3

DIRECTIONS: Show what is missing by drawing it. Use words to help make clear what your drawing means, if you wish. Use words instead of a drawing, if you can show that you know what is missing more easily and clearly by words. Each correct completion counts 2. Each wrong completion counts 1 off from your score.

(This test includes cuts of ten familiar objects, such as a spade, a saw, a ladder, a fish-hook, etc., from which are omitted essential parts, such as the handle of the spade, teeth of the saw, barb of the hook, etc.)

TEST 4

DIRECTIONS: In each line there are seven objects. In each line draw a circle around that one of the last four objects which has the same relation to the third object that the second has to the first. For example, if the line was head, hat, foot, hand, shoe, leg, glove, you would draw a circle around the shoe. Each line correct counts 2. Each wrong line counts 1 off.

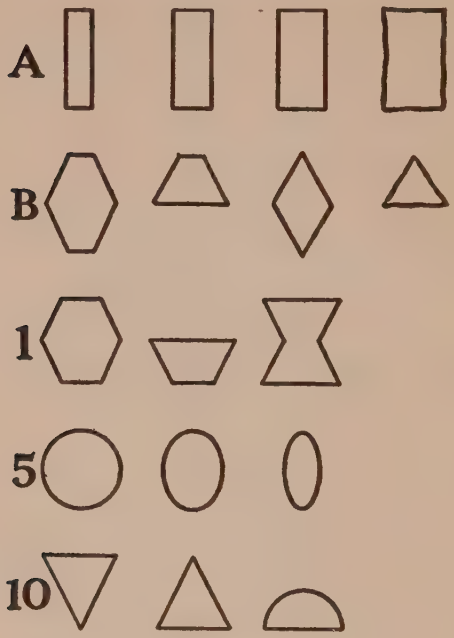
(Instead of words, cuts appear in the test.)

1.	Collar	collar button	cuffs	cuff links	shirt	stick pin	cravat
4.	Lighted candle	oil lamp	gas jet	matches	oil can	framed landscape	electric bulb

12. High wire fence dog window screen two flies hammer window sash door

TEST 5

DIRECTIONS: In lines 1 to 10 draw a fourth figure in each line such that the fourth figure is to the third as the second is to the first, as shown in lines A and B. Each line correct counts 2. A wrong line simply counts 0.



TEST 6

DIRECTIONS: Find the answers. Each correct example counts 5. There is no penalty for a wrong answer. It simply counts zero. Use the last page of this examination blank for your work if there is not enough space here.

1. $4 = a$. $5 = b$. $6 = c$. How much more is ab than 7—c? Ans. .....
4. $(x-4)^2 = x(x-3)$. What does x equal? Ans. .....
6. $c = \frac{x-a}{x-b}$. What does x equal? Ans. .....

TEST 7

DIRECTIONS: Answer as many of these questions as you can. You are not expected to know many. Each correct answer counts 4. Each wrong answer counts 2 off.

Women candidates should answer quickly any of these which they are sure they can answer correctly, and use the remaining time for this test in perfecting the work of previous tests.

1. What are the timbers called on which the flooring in a frame building rests? .....
5. What makes it possible for an automobile to turn a sharp corner and still keep the traction on both rear wheels? .....
7. What is the air gap of a generator? .....
10. How many flutes are there usually in a standard half-inch tap? .....

TEST 8

DIRECTIONS: Read statement 1. If it is true, write a capital T before the 1. If it is false, write a small f before the 1. If you do not know, go on to the next. Do the same with statement 2, and so on with each of the 60 statements. Each statement correctly marked counts 3; each statement wrongly marked counts 3 off.

1. If a cube is cut by a plane surface, the section will always be a square.
5. Sound is transmitted by transverse waves.
9. Hydrogen is lighter than air.
13. The outer whorl of members of the flower is the calyx.
17. The sea urchin is an invertebrate.
25. The magnetic axis of the earth points to the North Pole.
30. A protest of a note is a statement by the maker that payment is refused.
33. Music intended to illustrate some incident, story or poem is usually called popular music.
38. Ben Jonson wrote novels of politics.
41. The war with Mexico in 1845 was caused by the Mexican invasions of the United States.

44. France and England have always been allies.
49. Congress has provided no law determining the presidential succession in case of the death of both President and Vice-president.
54. It was Pericles who really made Greece supreme on the seas.
60. If the spark is too far advanced an engine will knock.

2. Instructions for Giving the Thorndike Intelligence Examination for High School Graduates

Series of 1919, 1920, 1921, and 1922

One examiner can take care of forty to fifty men, if he is quick in distributing the material.

The actual working time is:

10 minutes for fore-exercise with Part I, Practice Form.
30 " Part I, Trial 1.
30 " Part I, Trial 2.
60 " Part II.
40 " Part III.
2 hr. 50 minutes.

From twenty to thirty-five minutes is required for explanation, distribution of papers, and rest period. So three and one-half hours should be allowed.

The examiner should first read the following carefully, then put a friend or a few trustworthy students through the test as a genuine performance. If he makes any mistakes, he should repeat this. At the regular examination he should be sure thirty minutes before the time of beginning that he has the necessary material. This comprises for each candidate, for the series sent herewith:

2 pencils, sharpened.
1 Part I Practice Form blank with 13 tests.
1 Part I Form .. blank with 13 tests.
1 Part I Form .. blank with 13 tests.
1 Part II Form .. blank with 8 tests.
1 Part III Form .. blank with 3 tests.

and a copy of this set of Instructions for the Examiner and a watch that tells seconds.

Seat the candidates, provide each with two pencils. Say, "You are to take a test. All instructions necessary will be found on the test sheets themselves. Do not ask any questions of the examiner or of anybody. Do not look at any other person's paper. Use pen or pencil as you prefer. If you need another pencil at any time, come to the desk and get it."

Say, "This examination will comprise a preliminary inspection of the sort of thing you are to do, followed by four work periods,

two of a half hour each, one of an hour, and one of forty minutes. It is not expected that anybody will do correctly everything in the examination within the time allowed. There may be many things in the examination that you cannot do, but there will be some hundreds of different things in all. So there will be an abundant chance for everybody to show his ability. The examination is arranged so as to give the quickest person enough to keep him busy full time, consequently many of you will not complete all the work. A person may make a good score in the examination although he does not fully complete any division of it."

Then say, "You will be given a blank like this (holding up Part I, Practice Form). When you receive it read what it says on the outside page. Do not open it or do anything further until I tell you to."

Distribute Part I Practice Form quickly, giving the candidate at the end of each row enough to supply that row. Then say: "This first trial is not a part of the examination and will not influence your mark. It is for the purpose of giving you a notion of the examination you are to take."

At 00 say, "Turn the first leaf and examine Test 1. Do not write anything but think what you would write if you were really taking the test."

At 0.30 say, "Examine Test 2. Do not mark anything, but think how you would mark the sentences if you were really taking the test." And so on, at intervals of thirty and sixty seconds until, at 10.00 say, "Stop. Pass the papers to the right as quickly as you can." Collect the papers rapidly.

Then say, "You will be given an examination like that which you have just seen. You will have thirty minutes in all for it. When you receive the examination let it lie untouched until I say 'Go.' Do not open it or write until I say 'Go.' Read what it says on the outside page, and do what it says to do. When I say 'Go,' write your name, finish reading the instructions on the outside page, open to Test 1 and work as fast as you can. If your pencil breaks, get one from one of the boxes" (showing them where they are).

Distribute the blanks for Part I Form (....) quickly. Set your watch to a little before 0 min. 0 sec.

At 0 min. 0 sec. say, "Go."

At 1 min. 30 sec. say, "Even if you have not finished Test 1, begin Test 2"; and so on, at intervals of two or four minutes until, at 30 min. 0 sec. say, "Stop."

Collect all papers quickly, making sure that each is signed.

If a candidate hands in his paper before the thirty minutes is completed, look at it, especially at Nos. 3, 4, 8, and 12, to make sure that he has omitted no test. If he has not, enter on the front page the time at which he completed the test. Only about one in one hundred candidates will do all the work within the thirty minutes.

Say, "You will be given another examination like that which you have just done. Do not open it or touch it until I say, 'Go.' When

I say 'Go,' write your name, open it, and begin on Test 1." Distribute the blanks for Part I, Form (....) quickly, set your watch back to a little before 0 minutes.

At 0 min. 0 sec. say, "Go."

At 1 min. 30 sec. say, "Even if you have not finished Test 1, begin Test 2."

Follow exactly the same time schedule as in the first trial.

If a candidate hands in his paper before the thirty minutes is completed, look at it, especially at Nos. 3, 4, 8, and 12, to make sure that he has omitted no test. If he has omitted one, call his attention to the fact, returning his paper. If he has not, enter on the front page the time at which he completed the test. Only about one in one hundred candidates will do all the work within the thirty minutes.

Say, "You may have five minutes' intermission now, before the next test."

Ventilate the room thoroughly during the intermission. After five minutes, call the candidates to their seats, and say, "You will be given a blank like this (holding Part II up to show them) containing eight tests. When you receive it, read what it says on the front page and do what it says there, but do nothing further. Under no circumstances open the blank until the signal to begin is given. Extra credit for speed will not be given. So do not hand in your paper before time is called."

If those tested include women, add to the announcement of instructions for Part II the following:

"Some of the elements of this part of the examination are not well suited to women candidates. They are in Test 3 and Test 7. In Test 3 and Test 7, therefore, women candidates should do only what they are fairly sure they can do correctly, and should spend the time so saved in perfecting their work on the other tests on this blank."

Distribute the Part II blanks (eight tests) quickly. Set your watch a little before 0 min. 0 sec. When it is at 0 min. 0 sec. say, "Begin."

When it is at 8 min. 0 sec. exactly, say, "Even if you have not finished Test 1*a* begin Test 1*b*." And so on at intervals of 8, 8, 3, 3, 3, 10, 4, and 13 minutes until, when it is at 60 min. 0 sec. say, "Stop. Pencils down. Be sure your name is on the front of the blank." Have the papers collected promptly, seeing that each is properly signed.

Say, "The next part of the examination will contain three tests. You will have time enough to do careful, thorough work in this part of the examination. You will be given a blank like this (holding up Part III). When you receive it, read what it says on the front page and do what it says there. Do not open the blank until the signal to begin is given."

Distribute the Part III blanks quickly. Set your watch to a little before 0 min. 0 sec.

At 0 min. 0 sec. say, "Go."

At 10 min. 0 sec. say, "Even if you have not finished Test 1*a*, begin Test 1*b*."

At 20 min. 0 sec. say, "Even if you have not finished Test 1*b*, begin Test 2."

At 40 min. 0 sec. say, "Stop. Pencils down. Be sure your name is on the front of the blank."

Collect all the blanks quickly, making sure that each is signed. Dismiss the candidates, after making any announcements to them that have been authorized. Take all the examination blanks, including the Practice Form of Part I, to the proper authorities.

3. Instructions for Scoring the Thorndike Intelligence Examination for High School Graduates

Series of 1919, 1920, 1921, and 1922

GENERAL INSTRUCTIONS

Tests 1 to 13 of Part I (that is all of Part I), and Tests 4, 6, and 8 of Part II can be scored by any clerk. A person who could pass none of these could score them all perfectly. No judgment is required. Tests 3, 5, and 7 of Part II require only sufficient judgment to follow the key and seek advice from the chief examiner in cases that are not clearly wrong. Tests 1 and 2 of Part II, and 1*a*, 1*b*, and 2 of Part III do require ability of the grade of a student in the upper half of the college group. Great pains are, however, not at all necessary in scoring these tests. For the total scores from the forty-eight elements will be closely the same whether the scoring is rapid or scrupulously careful.

In Tests 1*a*, 1*b*, and 2 of Part II, and 1*a*, 1*b*, and 2 of Part III, the whole set of papers may best be scored for two or three elements, then for the next two or three, and so on. This method is in fact specified in the detailed instructions.

The routine suggested below for the place of entering the scores, and for the time and place of entering total scores in each test, is only one of many economical ones. It may be abandoned in favor of one better suited to special conditions.

No routine is prescribed for entering the total score for Part I, Trial 1, Part I, Trial 2, Part II and Part III, finding the sum of the four and entering it. This may be done in any way that best fits the system of cards or books used. It is, however, suggested that the final record of the examination should include not only the grand total sum, but also the four separate totals for Part I, first trial, Part I, second trial, Part II and Part III. Further, if the candidate's actual papers are not kept, it may be desirable to keep a record of his scores in the single tests, so as to have the means to measure him separately in the mathematical, verbal, pictorial, and informational tests.

With an adequate staff and proper supervision, the number of papers that can be scored in a day is practically without limit. The work of each scorer should be specialized.

INSTRUCTIONS FOR SCORING PART I (13 TESTS)

(Use blue or red pencil in all scoring)

Each numbered task under a test will be called an element. Thus, Test 1 has 5 elements, Test 2 has 10 elements, Test 3 has 8 elements, and so on:

The score is to be:

Test 1. Number of correct elements.
" 2. Number of correct elements minus number of wrong elements.
" 3. Number of correct elements.
" 4. Number of correct elements multiplied by 2.
" 5. Number of correct elements.
" 6. Number of correct elements minus number of wrong elements.
" 7. Number of correct elements.
" 8. Number of correct elements.
" 9. Number of correct elements.
" 10. Number of correct elements. Count a column correct only
" if BOTH markings are right.
" 11. Number of correct elements.
" 12. Number of correct elements minus number of wrong elements, but give no score lower than 0.
" 13. Number of correct elements.

The maximum obtainable is thus 153 or 154 (5, 10, 8, 20, 10, 20, 5, 10, 20, 9, 8, 8 or 9, 20).

INSTRUCTIONS FOR SCORING PART II (8 TESTS)

(Use blue or red pencil in all scoring.)

(The instructions are given in too great detail for reproduction here. They consist mainly in describing the proper use of the Keys that are supplied with the examination.)

INSTRUCTIONS FOR SCORING PART III

Take the key for Test 1*a*. Score 3, 2, 1, 0, —1, or —2 according to the instructions for scoring 1*a* in Part II. At the bottom of the page enter as a total score the algebraic sum of the credits obtained.

Take the key for Test 1*b*. Score in the same way.

Take the Key for Test 2. Score 3, 1, 0, or −2 according to the instructions for scoring Test 2 in Part II. At the bottom of the page enter the algebraic sum of credits.

As the total score for Part III, use one and one-half times the sum of the three entries at the bottom of pages 2, 3, and 4.

INSTRUCTIONS FOR COMPUTING THE CANDIDATE'S TOTAL SCORE

Add the scores in Part I first trial, Part I second trial, Part II, and Part III. Multiply this total by 0.2.

The scores computed in accordance with the instructions issued will vary from about 40 to 100, with an occasional record above 100. Individuals scoring under 60 who are seventeen years old or older are as a rule unsuitable material for college education. If under seventeen, they may be advised to study or work for a year and then submit to a second examination. Individuals scoring from 60 to 69 have, probably, enough intellect to attain a college degree, if they are specially earnest and industrious. Their admission to college may be determined by considerations of general or special policy. Individuals scoring 70 or over possess intellect such as is adequate for college work and may be admitted with no risk of detriment to present standards. Individuals scoring 85 or over may be safely admitted regardless of even gross deficiencies in their preparation.

These statements concern candidates who have had good educational advantages, and who know English as a mother tongue. The expectation from equally intelligent graduates of poorly equipped high schools and from equally intelligent foreign students will be less. These statements also refer to courses in American colleges and engineering schools of high standards. For many institutions lower figures would be more appropriate. Further information concerning the significance of the scores will be furnished upon request.

APPENDIX TWO

1. Data on the Separate Parts of the Thorndike Examination

Table 54 shows the distributions of the scores obtained on each Part by the 505 candidates for admission to Columbia

TABLE 54

Score	I D	I E	II B	III B
0-9			1	1
10-19			1	2
20			1	7
30			4	36
40			7	67
50	2	1	5	111
60	9	1	14	111
70	13	5	25	88
80	27	12	29	55
90	77	49	28	17
100	88	94	50	10
110	111	136	50	
120	94	126	51	
130	66	69	56	
140	18	12	49	
150			36	
160			33	
170			29	
180			19	
190			10	
200			4	
210			2	
220			1	
N	505	505	505	505
M	111	115	124	62.5
Sigma	17.8	14.3	37.2	17.2

Table 54. Distributions of scores on the Parts of the Thorndike Examination of 505 candidates for admission to Columbia College who took the Examination in September, 1919. The first frequency column represents the scores on the first half of Part I, and is headed I D. The second column is for the second half of Part I (Form E). The third is for Part II, Form B. The fourth is for Part III, Form B.

TABLE 55

Score	I D	I E	II B	III B
25				1
30				
35				1
40				5
45				3
50				6
55				20
60				13
65			2	13
70				17
75	2		3	10
80	1			6
85	3		3	9
90	6	4		2
95	10	1	4	3
100	5	8		1
105	10	10	19	1
110	10	14		
115	15	19	9	
120	14	17		
125	14	9	11	
130	8	15		
135	8	11	14	
140	3	2		
145	1	1	14	
150	1			
155			1	
160				
165			12	
170				
175			7	
180				
185			6	
190				
195			5	
200				
205			1	
N	111	111	111	111
M	114	118	130	66
Sigma	15.9	12.5	32.6	14.6

TABLE 55. Distributions of scores on the Parts of the Thorndike Examination of 111 New Plan men admitted to Columbia College in September, 1919, who were able to stay in college two full years. Each distribution in this table may be compared with the corresponding and similarly named distribution in Table 5.

The score-index is the midpoint in each respective step.

College who took the examination in September, 1919. Of these men, 312 were New Plan candidates, not all of whom were admitted. Exclusive of foreigners, about 111 of the New Plan candidates admitted in September, 1919, were still in Columbia College at the end of the second year. Table 55 shows how the scores on the Parts of the examination are distributed for these 111 men.

2. Distributions of Scores of Various Groups of Tests in the Thorndike Examination

In the following distributions the most significant features to notice are the differences in variability and in type of curve exhibited by these groups of tests. The scores on these groups are summated without alteration in finding the Total Score on the Examination, and therefore the weighting of each group in the Total Score is directly measured by the S. D. of each group.

The Language Group is composed of Tests 1, 2, 6, and 11 of Part I. The time required for this group is sixteen minutes; the S. D. is only 8, with Mean at 70; and the skewness indicates that the group of tests is too easy for high school graduates, since the scores are not well distributed at the upper end.

The Mathematical Group is composed of the arithmetic fundamentals, arithmetical reasoning, the number-series completion tests of Part I, and the algebra test (6) of Part II. This group, requiring twenty-six minutes, of which ten minutes are for the algebra test of Part II, gives an excellent distribution. It contrasts so strikingly with the Language Group as to form of curve and variability as to suggest the revision of some of the tests so as to make them difficult enough to differentiate all along the line from dull to very bright high school graduates. In general, no test should be so easy that an appreciable percentage of examinees get perfect scores. A more detailed discussion of this point will be found in Appendix III, 2.

TABLE 56

Score	Language I,1,2,6,11	Math. I,3,4,8, II,6	Informa. I,5,II,8	Reading II,1+III,1	Completion II,2+III,2	Logic I,9+12	Non-V I,10,13, II,3,4,5	Trade II,7	
								Score	Freq.
−10			1					−6	5
−5								−4	5
0				1				−2	18
5			3		6			0	16
10			2	1	8	2		2	14
15		1	5	3	24	10		4	18
20			3	1	27	14		6	10
25		2	12	9	32	16	1	8	17
30		3	9	23	31	23		10	31
35	1	5	7	33	22	43		12	21
40		8	15	32	17	38		14	7
45	2	13	17	32	13	21		16	7
50	6	12	27	30	3	12	3	18	5
55	6	16	20	18	3	9	2	20	4
60	14	22	18	5	1		3	22	5
65	28	21	15		1		7	24	2
70	53	22	9				11	26	2
75	54	14	6				27	28	1
80	19	13	8				30		

85	5	16	3				28		
90		13	3				22		
95		3	2				30		
100		3	2				34		
105		1	1				90		
N	188	188	188	188	188	188	207 [1]		188
M	69.8	65.2	51.	40.7	30.5	35.1	85.4		7.39
Sigma	8.2	16.9	20.	10.2	23.2	10.3	13.		7.34
Time in Minutes	16.	26.	17.	36.	28.	8.	21.		4.

TABLE 56. Distributions of scores of groups of tests in the Thorndike Examination. These represent the scores of the men admitted under the New Plan in September, 1919, who survived in Columbia College one year. The index-score is the midpoint of each respective step. Note that the distribution of the Trade test scores (II, 7) is on a scale of steps different from that of the other distributions in this Table.

The legends of the distributions are to be interpreted as follows: the Roman numerals indicate Parts of the examination; the Arabic numerals indicate tests in the specified Parts. Thus "I, 1, 2, 6, 11," is to be read, "Part I, Tests 1, 2, 6, and 11." It should be remembered that Part I consists of two Forms, and therefore "I, 1, 2, 6, 11" really represents two tests of each of four kinds.

1 This distribution is taken from the records of students admitted in September, 1920, who stayed in Columbia one year. There is no reason why the corresponding distribution for the September, 1919, men should be different.

3. Stability of the Means and S. D.'s of the Parts of the Thorndike Examination

The data presented in Table 56 may be analyzed so as to resolve the same question applied to the various Parts of the examination. As has already been noticed, the two Forms were made up as follows:

	Part I		Part II	Part III
1919	C	D	B	B
1920	E	F	C	C

Distributions of scores for each of these eight parts are presented in Table 57.

The Means vary considerably; but the most important feature of these data is the comparative stability of the S. D.'s. The Means may vary a good deal without vitiating the examination: all that we need is to know that they do vary and how much. Of course, these facts were known to the author of the examination as exactly as experimental methods would allow; and it was this knowledge which enabled him to secure almost perfect stability in the Mean of the total examination. It is also possible to correct for differences in variability, but the procedure is complex and somewhat laborious; thanks to the skill of the author of these tests, this labor is spared us.

4. Weighting of the Parts of the Thorndike Examination

The question arises at this point whether the varied tests and Parts of the Thorndike battery are weighted in the best possible way for the special purpose of the prediction of general college success as defined by our Scholarship Score. I have nothing beyond a few indefinite inferences to offer on this point with regard to the individual tests; and these will be found in Chapters Two, Seven, and Eight. But the weight-

TABLE 57

Score	I C	I E	I D	I F	II B	II C	III B	III C
−5							1	3
0						1		3
5					1			4
10								10
15							6	17
20					2			24
25					3		21	21
30					4			29
35					10	3	36	17
40					4			10
45					8	2	44	13
50	2				11	4		7
55	1	1			10	8	40	4
60	1				8	10		6
65	3		1		11	5	14	1
70	4	3	1	1	16	8		2
75	15	4	2		14	17	6	
80	14	11	9	5	5	11		
85	14	10	6	6	9	18	2	
90	15	15	12	13	7	11		
95	21	24	21	19	7	13	1	
100	27	21	22	16	11	8		
105	18	28	18	22	3	8		
110	11	23	27	22	6	8		
115	14	18	20	28	4	6		
120	6	3	18	16	4	7		
125	3	5	7	13	1	5		
130	1	3	2	8	4	4		
135	1		2	1	2	3		
140			3	1	1	4		
145					1	2		
150					2	1		
155					1	2		
160					1			
165								
175								
N	171	171	171	171	171	171	171	171
M	95.3	100.	105.	107.	75.	90.	40.	29.
Sigma	15.8	13.5	14.2	13.5	30.3	27.8	15.4	15.2

TABLE 57. Distributions of scores on each Part of two Forms of the Thorndike Examination given to 171 students in an Eastern State Normal School. One Form was given in 1919, the other one year later. About ninety per cent of these students were women. The score-index is the midpoint of each respective step.

ings of the Parts have been studied with practical exactness by the use of three different sets of data. The results of the three determinations agree closely in indicating that the weights assigned by Professor Thorndike are the best possible weights.

The best method of determining the proper weights for the Parts is to calculate the partial regression equation in which C, the criterion, would be the Scholarship Score, and x_1, x_2, x_3, and x_4 would be Parts I*a*, I*b*, II, and III. This method involves the solution of a five-variable determinant for each regression equation—a procedure which is quite laborious, and

TABLE 58

	I	II	III	IV	V	M	Sigma
I. Freshman Year Scholarship Score	1.00	.58	.47	.50	.482	220	52.8
II. Part II	.58	1.00	.46	.585	.59	135	32.60
III. Part III	.47	.46	1.00	.54	.545	63	14.60
IV. Part I*b*	.50	.585	.54	1.00	.70	115	12.50
V. Part I*a*	.482	.59	.545	.70	1.00	115	15.90
Weights		1.00	.5585	.4157	.1529		
Total Correlations		.58	.6235	.6351	.6362		

which happily has just been rendered unnecessary for our purpose by a discovery of Dr. H. A. Toops of the Institute of Educational Research, Teachers College, Columbia University. Dr. Toops' method, for which the basic formulæ were derived and published by Dr. T. L. Kelley several years ago, does not give us *the* regression equation, as we have become used to thinking of that term, but it does give us *one* of many possible equations that satisfy the conditions which were supposedly fulfilled only by the strict regression equation based upon the Partial Correlation method mentioned above. The essential difference between the Toops equations and the strict regression equation is that in the former the weights are determined in series, one at a time as each new variable is

added to the battery. As a consequence, the weightings depend in part upon the order in which the variables are added, but the maximum multiple correlation is practically the same for all the equations derived by Dr. Toops' method or by the strict partial correlation procedure. The advantage of Toops' equations is that the labor and time for calculating them is a very small fraction of that required for the solution of a determinant, and this fraction decreases rapidly as the number of variables increases.

The determinant of Table 58 is based on the Freshman-year records of 111 New Plan men admitted in September, 1919,

TABLE 59

	I	II	III	IV	V	M	Sigma
I. Sophomore Scholarship Score	1.00	.545	.452	.451	.454	220	65.00
II. Part II	.545	1.00	.46	.585	.59		
III. Part III	.452	.46	1.00	.54	.545		
IV. Part Ib	.541	.584	.54	1.00	.70		
V. Part Ia	.454	.59	.545	.70	1.00		
Weights		1.00	.5971	.9028	.0615		
Total Correlations		.545	.59	.6241	.6244		

who remained in Columbia College two years. The correlation coefficients are exact only to two figures. The maximum correlation between the criterion and the best possible combination of the Parts according to Toops' formula is 0.6362. Allowing for the fact that the raw correlations in the determinant are exact only to two places, we may concede that this coefficient is really 0.64. Now if this were significantly greater than the correlation given by Thorndike's weightings, we should naturally advise the abandonment of the latter in favor of Toops' weights. Table 27 shows the maximum correlation given by Thorndike's weights to be 0.632. The difference is 0.008 in favor of Toops' equation—a difference so small as to be negligible. The experience and insight of Professor Thorn-

dike enabled him to weight the Parts of his examination so advantageously that a careful mathematical study of empirical results can increase the total correlation only by an amount not in excess of 0.008.

Table 59 is a determinant based on Sophomore-year records of the same 111 men as were involved in Table 58. Here we see the results of Table 58 verified. The increase in total correlation effected by weights empirically determined by Toops' equation is so small as to be negligible. Thorndike's weights give $r = 0.620$, and Toops' weights give $r = 0.6244$.

TABLE 60

	I	II	III	IV	V	M	Sigma
I. Freshman and Sophomore Scholarship	1.00	.60	.49	.446	.555	440	109.00
II. Part II	.60	1.00	.46	.59	.585		
III. Part III	.49	.46	1.00	.545	.54		
IV. Part I*a*	.446	.59	.545	1.00	.70		
V. Part I*b*	.555	.585	.54	.70	1.00		
Weights		1.00	.5712	.0688	.6231		
Total Correlations		.60	.6466	.647	.6675		

An interesting thing to observe in Tables 58, 59, and 60 is the consistently small increment in total correlation made by the addition of Part I*a*. This Part adds at least forty minutes to the examination time and constitutes probably fifteen per cent of the total cost, including scoring, recording, etc., of the examination.

Table 60 only confirms the conclusions derived from the two next preceding Tables. This determinant is based on the combined Freshman and Sophomore records of the 111 men of Tables 58 and 59. The total correlation given by Toops' equation is here slightly in defect of the empirical correlation given by Thorndike's weights and noted in Table 22. This is undoubtedly due to the slight inexactness introduced by using

raw correlations correct only to two decimal places in the determinant, and possibly to some slight chance factor which may have raised the obtained correlation above the true correlation by a small amount.

APPENDIX THREE

Suggested Improvements in the Thorndike Examination

1. Each separate test in the examination should have a definite and fixed time allowance. This suggestion is made in the interests of reliability and analysis. With the present method of administering the examination, no effort is made to induce all examinees to spend the same amount of time on each test. The theory of this variable distribution of time is that each candidate is permitted to spend his time on those particular tasks which he finds easy to perform, and which will therefore net him the highest number of credits. But it is not certain that the candidate is capable of judging what tasks in the examination he is best able to cope with, even if he were perfectly familiar with the weights of the various tests. The improvement of the examination depends at least partially upon our ability to learn the significance of each type of test in it. Such analysis is at present impossible. For example, the real significance of Test 6 in Part II may be three times as great as that of Test 5 in Part II, but due to some fortuitous circumstance in the superficial appearance of Test 6 so many candidates might avoid it who could (if they tried) get high scores on it that its correlation with the criterion might be reduced spuriously below that of Test 5. The capricious distribution of time allowed under the present organization of the examination may also decrease the reliability and significance of the whole examination. Even if each candidate were infallible in choosing the tasks which best fitted his peculiar talents and counted most heavily in the final score, it is a question whether he ought to be allowed to do so.

2. Each test should include elements easy enough for the least intelligent examinee to solve successfully, and elements difficult enough to keep the most intelligent examinee busy up to the time-limit. A glance at the distributions in Table 36, page 95, will show the propriety of this suggestion. Much of the value of Tests 1, 2, and 7 of Part I is lost because they are so easy that they do not differentiate among the upper fifty or sixty of every hundred boys. In general we may say

that the ideal test, other things being equal, is one which fulfils the following conditions:

(*a*) The Mean should approximate fifty per cent of the possible maximum score;

(*b*) The distribution should be fairly smooth and symmetrical, i.e., the increments of difficulty should be both small and uniform;

(*c*) Extreme scores of a distribution should neither exhaust nor fall too far short of the limits of the test above or below.

3. The Reading paragraphs should be increased in number and put in a separate section.

The Thorndike Reading Test has so many strong points in its favor that the strengthening of it is a manifest duty. Its one limitation is its unreliability. The three facts upon which this suggestion is based are:

(*a*) The Reading Test has never failed to correlate with an acceptable criterion up to the limit set by its reliability coefficient;

(*b*) It has been experimentally demonstrated that its reliability increases very nearly in accordance with the indications of Brown's formula; and

(*c*) It is an ideal test in that it is non-coachable, since the number of paragraphs that can be made up is practically limitless; and all effective attempts to prepare for it would result in the achievement of a legitimate and very valuable educational objective—the ability to read English prose with incisive understanding.

It is suggested that at least six, and preferably eight, paragraphs be included; and that these eight Reading paragraphs be made into a new Part III. This would automatically stabilize the time allowance for the Reading Test, and would effect other conveniences in administration and scoring.

4. The number of elements in the Completion and Information Tests should be increased.

These two Tests uniformly give correlations with the criterion up to the limit set by their reliability coefficients; we have abundant evidence that their reliabilities can be improved by increasing the number of elements; and they have the same ideal qualities as the Reading Test, in that they are non-coachable, and any effective attempt to prepare for them would result in genuine educational advancement.

The Completion Test now consists of two groups of twelve sentences each. Twenty-four sentences are not enough:

(*a*) to give an equitable sampling of material and thereby reduce the very large influence which specialized information and experience have in this type of test,

(*b*) to give an adequate finely-graded sampling of performances in a very complex activity, and

(*c*) to counteract the unavoidable subjectivity and dependent irregularities in the scoring of "completed" sentences. Forty-eight would seem a fair number of sentences for a reliable Completion test of intelligence, while forty is certainly the absolute minimum below which we cannot afford to go.

The same argument applies to the Information Test. The wider the sampling of material, the less it becomes a test of *special* information and the more it becomes a test of capacity and judgment. At the same time as we increase the number of elements we decrease the influence of all chance factors. The test now consists of ten Recognition and sixty True-False questions. These forms of questions are so easy to get and so efficient in administration and scoring that it would seem reasonable to raise their combined number to 125 at least. By increasing the number of True-False questions in Test 8 of Part II it would become possible to secure a fairly reliable "content" index for each candidate. In its present form, Test 8 of Part II is practically an examination in secondary school contents, and it alone probably gives as reliable a measure of high school achievement as the High School marks give. The information test thus presents an opportunity for the intelligence test to combine with itself a reliable measure of the candidate's preparation. We have already said in preceding chapters that both intelligence and content examinations would find their greatest serviceability when used together. Since effective improvement in the traditional content examinations may be delayed for some time, would it not be advantageous for the intelligence examination to assume some responsibility *ad interim* for supplying our admissions offices with objective, comparable and reliable measures of secondary school content?

5. The Opposites Test should be lengthened by the addition of more difficult elements.

The great value of this test in the measurement of intellect

or native capacity is an accepted fact which rests upon twenty years of experimentation. But much of that value is lost in the Thorndike Examination by reason of the shortness and ease of the test. About eighty per cent of Columbia College entrants secure the maximum possible credit on the Opposites Test (I, 6). By increasing the length and difficulty of this test, the differences that exist among the eighty per cent who now get the same score would not be so effectively concealed as at present.

6. The Mathematical Tests (I, 3 Arithmetic Operations, I, 4 Arithmetic Reasoning, I, 8 Number Series Completion, and II, 6 Algebra) might be increased to advantage.

The significance and reliability of this group of tests are so great that no effort ought to be spared to increase them. The reliability is now above 0.80, and the significance above 0.65. It seems probable that by increasing the fineness in the gradations of difficulties in these tests, and increasing the time-allowance slightly, the reliability of the group might rise above 0.90, with a corresponding rise in significance.

7. The "fore-exercise" should be made more effective.

The purposes of the fore-exercises are,

(*a*) To allay nervousness,

(*b*) To decrease the possible influence of unfair coaching,

(*c*) To acquaint the student with the "general idea" of the examination so as to decrease the unreliability due to poor understanding of directions and to loss of time spent by the uninitiated in amazement, awe, and wonder in the face of tasks of a very puzzling aspect.

By the testimony of many Columbia College entrants who took the examination, the fore-exercise fails to allay nervousness. In some cases it positively aggravated nervousness by giving an exaggerated impression of the necessity for haste. The present time-limits are not adequate for acquainting the uninitiated with the "general idea" or with the directions sufficiently well to neutralize any unfair coaching on the part of isolated individuals. The purposes of the fore-exercise as stated above could be achieved almost without exception

(*a*) by giving out copies of the fore-exercise a day or two (or several hours) before the examination; or

(*b*) by allowing twenty or thirty minutes instead of only ten for perusal of the fore-exercise at the examination.

8. The elimination of one-half of Part I is recommended. This seems desirable because:

(*a*) The two forms of Part I correlate one with another very highly;

(*b*) There is little difference in total correlation with the criterion when the whole or only one-half of Part I is used;

(*c*) The use of more time on the fore-exercise would probably add as much to the reliability of one-half of Part I as the use of the other half would add;

(*d*) Its elimination would save at least thirty-five minutes of very precious time in the administration of the examination, thus making it easier to keep the time for the total examination within the crucial limits of three hours.

(*e*) Its elimination would reduce the cost of the examination blanks by approximately one-fifth, and would reduce the time and cost of scoring by nearly one-fifth.

9. The elimination of Tests 10 and 13 of Part I and of Tests 3 and 4 of Part II is recommended, independently of, and in addition to, all the other recommendations made in this chapter.

Test 10 of Part I is not a test of intelligence for the candidates who seek admission to Columbia College. Four different measures of its significance give it a maximum correlation with college success of $\pm$ 0.05. Ninety-six per cent of its scores are between fifteen and eighteen, inclusive, eighteen being the maximum possible score.

Test 13 of Part I correlates with the criterion to an extent not greater than 0.20. This is not insignificant, even though at the expense of six pages in the examination blank and eight minutes of the examination time; but its exclusion is recommended as a means of making room for more significant types of tests.

Test 3 of Part II gives a negative correlation, $-$ 0.14, with the criterion, and probably contributes an appreciable part of the unreliability of the whole examination. It is a test of a very specialized type of information in a particularly limited and inefficient form of question. The pictorial method of testing is costly even among the illiterate where it is necessary; its use among literate candidates of as high an order of intelligence as High School graduates is a doubtful economy. There

are more efficient means than pictures for testing experience in the field covered by this test.

Test 4 is similarly a picture test, with only eight chances to win or lose. It gives a truncated distribution showing that it is too easy a task to differentiate among High School graduates; and correlates with college success less than 0.10.

10. The modification of Tests 1, 2, 7, 11, and 12 of Part I and of Test 7 of Part II is recommended.

In spite of its brevity and truncated distribution, Test 1 of Part I gives a correlation of 0.25 with the criterion. The Test should be made longer, but with only a small increase in time allowance.

Test 2 of Part I gives a near-zero correlation with the criterion, and its retention is of doubtful wisdom. But if it is retained, it ought to be made longer and put on a fixed time allowance not greater than is now given to it. Ten True-False statements scored by the Right-Minus-Wrong Method constitute a test which is necessarily highly unreliable.

Test 7 of Part I consists of five very easy elements, gives a very steep truncated distribution, and a correlation with the criterion very near zero. In its present form the test is not worth including; but if the elements were increased to at least twenty and made much more difficult, and the present time allowance retained, it would very likely be a valuable part of the examination.

Test 11 of Part I is too short, and also suffers by having ambiguous directions for the giving and scoring of it. The directions to the examinee say "write the letter F before each of these statements which could not possibly be true." The result is that it is impossible to know which statements the examinee has omitted. This could be corrected by directing the candidate to put some other sign before those which he thought were not necessarily false. The directions for scoring say that those statements below the last one which the candidate has tried should be ignored. But it is impossible to tell which was the last question *tried* by the candidate.

Test 12 of Part I is too inelastic. There ought to be more and smaller units.

The ten Trade-form questions of Test 7 in Part II give correlations with the criterion very near to zero. In addition to the fault of inadequate sampling, these questions introduce

a large unreliability factor into the Total Examination Score by reason of the Draconian penalties which are visited upon those who make wrong responses. The student does not know that the questions relate to very technical experiences and that the words are used in an extraordinary sense. "What holds a horse on the ground?" The High School lad who has just been studying physics naturally answers "Gravity," and thereby has a certain number of points subtracted from his score. The most legitimate use of negative scores is in countering chance influences that are fairly measurable as in the True-False form of test. It is difficult to justify negative scores on any other grounds.

11. It is recommended that the scoring be simplified wherever possible. The general impression that users of the Thorndike Intelligence Examination get is that it is a very complex piece of machinery. Only a few of the tests have a common scoring procedure. Not only are the elements weighted variously in the different tests, but the procedure for applying these weights is quite variable. These mechanical details are quite trivial in comparison with sound general principles of mental measurement, but the administrative aspects of measuring instruments cannot be neglected with impunity.

(*a*) Wherever possible the responses should be in a straight line or column. It is a small matter, perhaps, but it saves time, expense, and often serious errors. In Test 3 of Part I the problems could have been placed in a vertical column on the right side of the page, with dotted lines for the answers also in a vertical column, and all the left three-quarters of the page available for the calculations of the examinees. In Tests 5 and 9 of Part I the scoring could be done with half the present time and expense and effort if the alternatives had been numbered and a space at the left of each statement provided for the response:

.... Gas engines are lubricated by (1) gasoline, (2) air, (3) water, (4) oil.
.... The color of oxygen is (1) violet, (2) greenish, (3) white, (4) brown, (5) pale blue, (6) colorless.

This important and useful plan was devised by Dr. Otis and

first used in the Otis *General Intelligence Examination* and the Otis *Self-Administering Tests of Mental Ability*. In Test 13 of Part I all the sketches on the last two pages should be numbered, thus permitting the identities to be indicated by putting the number of the elected sketch in a space provided for the purpose opposite each of the twenty elements on page 1 of Test 13. These twenty elements could have been printed in a vertical column without losing the necessary groupings by fives.

(*b*) The responses should be as brief, as easy to write and as easy to read as is consistent with their purpose. The scoring of Test 8 in Part II would be easier, quicker, more accurate and more agreeable if numbers were used instead of the plus- and minus-signs or the letters *t* and *f*. The numbers 2 and 4 have a strong suggestion of *true* and *false* in their sounds, and present a strong contrast in general form to the eye. These remarks apply equally well to Test 2 of Part I (Disarranged Sentences). In general, numbers are better than underlinings or signs of any other sort.

(*c*) Each response should be as definitely located as possible. In Test 8 of Part II a short dotted line should be provided at the left of the number of each statement. This small device adds nothing to the cost of printing, and saves the candidate the distraction of deciding just where the response should be recorded, saves the scorer from the annoying task of searching for each response, and lessens the probability of error.

(*d*) Space should be provided for the checks or marks and calculations of the scorer, and these marks and calculations should be as invariable as the variety of tests permits, and thoroughly systematized. Clerks will frequently have to be shifted from one position to another in the scoring process, and they will be called upon to interpret each other's records. Much confusion and many mistakes will be avoided if every clerk uses exactly the same system of marking and weighting each and every test in the battery. There is no one *best* way of scoring, but it is better to have *one* and *only one* way which is at least passable, than to have *many* best ways all in use at the same time on the same group of examinations.

It is nearly always best to have one thing done at a time in the objectively and mechanically scored tests. The weight-

ing of the tests should be done in the same way throughout the whole examination, and should be an operation separate and distinct from all other operations in the process of scoring. Two methods, illustrated in the scoring of Tests 6 and 8 of Part II prevail in the Thorndike Examination. In Test 6 the responses are each credited 0 or 5, and the summation of the 5's is the score on the test. In Test 8 each response is credited 1 or — 1, and the difference between the sums of the 1's and — 1's, multiplied by three, is the score on Test 8. The latter is the better method, and should be used systematically in every objectively scored test in the examination.

The first systematic act of the scorer should be to check, in a defined and regular way, each correct response; if the incorrect responses are desired, the next step is to check them in a defined and regular way; the next step is to count the right responses, and then the wrongs, if these have been checked. At the end of each test one of the following four schemes should appear:

(1) Number Right ..... minus Number Wrong ..... = ... = Score.
(2) Rights ..... minus Wrongs ..... = ... x W = ... = Score.
(3) Number Right ..... = Score.
(4) Number Right ..... x W = ... = Score.

All additions and multiplications should be checked at least once, and some at least twice.

In so far as this discussion does not conflict with the assignment of partial credits, as in the Reading and Completion Tests, it applies with equal force in the case of the subjectively scored tests. In the Reading and Completion Tests, where the algebraic sum of the credits assigned to the various elements is the score, it would undoubtedly be worth while to provide spaces for the sums of the positive and negative credits, both because it would lessen the number of petty mistakes made in the addition of the credits (even where they are all of the same sign, small errors are frequent) and because the availability of the negative scores on every test would facilitate exact studies of their relations to intelligence and of their influence on the Total Examination Score. Ob-

jective data on negative scores would be highly welcome. After each Reading and Completion Test, the following scheme should appear:

Credits ..... minus Penalties ..... = ... = Score.

And if, as in Part III of the current Series, the sum of the scores of n Reading and of N Completion Tests are to be weighted by a factor W, the following scheme is suggested as a preventive of mistakes:

Score of Reading Test 1.....
2.....
.....
.....
n.....
Sum of Reading Scores..... .....

Score of Completion Test 1.....
2.....
.....
N.....
Sum of Completion Scores..... .....

Sum of Reading and Completion Scores ..... x W = ... = Weighted Score.

In order to give a concrete picture of the revision which the writer would suggest as a result of three years of study specifically on the problem of the measurement of college intelligence, the skeleton outline of a complete College Entrance Intelligence Examination in three Parts, with the types of tests, number of elements in each test, and time allowances indicated, is shown on the following page. In order that the nature and extent of the changes recommended may be easily sensed, the skeleton of the Series of 1919-1922 is placed at the left.

THORNDIKE SERIES OF 1919-1922			SUGGESTED REVISION		
Test	Elements	Time	Test	Elements	Time
Fore-exercise, a Form of Part I.			Fore-exercise, a special blank containing samples of the fifteen different kinds of tests included in the revised examination.		
PART I					
1. Directions	5	1.5	1. Directions	15	3
2. Disar. Sentence	10	2.5	2. Disar. Sentences	25	3
3. Arith. Operations	8	2.	3. Arith. Operations	20	4
4. Arith. Reasoning	10	4.	4. Arith. Reasoning	25	8
5. Information	10	2.	5. Information	40	4
6. Opposites	20	2.	6. Opposites	40	3
7. Judgment	5	2.	7. Judgment	16	3
8. No. Series Com.	10	2.	8. No. Series Compl.	25	4
9. Analogies	20	2.	9. Analogies	40	3
10. No. Checking	9	2.	10. Absurdities	20	3
11. Absurdities	8	2.	11. Logic (syllogisms)	20	4
12. Logic (syllogisms)	9	2.		286	42
13. Forms Recogn.	20	4.			
	144	30.			
1 to 13 repeated	144	30.			
	288	1 Hr.			
PART II					
1a. Reading Par.	6	8	1. Completions	10	6
1b. Reading Par.	6	8	2. "	10	6
2. Completions	12	8	3. "	10	6
3. Pict. Compl.	10	3	4. "	10	7
4. Pict. Analogs.	8	3	5. Algebra	15	15
5. Geom. Forms	10	3	6. Information	100	20
6. Algebra	5	10		155	1 Hr.
7. Information	10	4			
8. Information	60	13			
	127	1 Hr.			
PART III					
1a. Reading Par.	6	10	1. Reading Paragraph	6	7
1b. Reading Par.	6	10	2. " "	6	7
2. Completions	12	20	3. " "	6	7
	24	40	4. " "	6	7
			5. " "	6	8
			6. " "	6	8
			7. " "	6	8
			8. " "	6	8
				48	1 Hr.
Fore-exercise and four separate blanks; total working time, 160 minutes; total elements, 439.			Fore-exercise and four separate blanks; total working time, 162 minutes; total elements, 489.		

BIBLIOGRAPHY

No pretense of completeness is made in including this short list of books and articles on the history, principles, and use of intelligence and educational tests. The effort has been made to restrict the references as much as possible to the field of tests in institutions of higher learning; but some books and articles of more general appeal have necessarily been included. Among these general references the writer hopes that the reader will not overlook those on statistical methods. It is needless to say that no one can cultivate an interest in mental tests to maximum advantage who neglects to learn the principles of the application of quantitative methods to education.

Complete up-to-date bibliographies will be issued from time to time by the National Research Council through its recently created Committee on College Entrance Tests. For further information, address the Chairman of the Committee on College Entrance Tests, National Research Council, 1701 Massachusetts Avenue, Washington, D. C.

The references below are arranged in alphabetic order of author's names:

Bates, Lt. Col. R. L. "A Study in Grades and Grading under a Military System." *Journal of Experimental Psychology*, Vol. 5, No. 5, page 329. October, 1922.

Breed, F. S. "Shall We Classify Pupils by Intelligence Tests?" *School and Society*, Vol. 15, pages 406-409. April 15, 1922. (A very suggestive summary view of the problem in the large, with interesting recommendations.)

Breitwieser, J. V. "The Thorndike College Entrance Tests in the University of California." University of California Press, Berkeley, California. June, 1922.

Breed, F. S. "Shall We Classify Pupils by Intelligence Tests?" Houghton Mifflin Company, Boston. 1922.

Burtt, H. E., and Arps, G. F. "Correlation of Alpha Intelligence Test with Academic Grades in High Schools and Military Academies." *Journal of Applied Psychology*, Vol. 4, pages 289-293. 1920.

Chambers, G. G. "Intelligence Examinations and Admission to College." *Educational Review*, Vol. 61, pages 128-237. 1921.

——— "Intelligence Examinations and Admission to College." *Educational Review*. February, 1921.

Colvin, Stephen S. "Educational Guidance and Tests in College." *Journal of Applied Psychology*, Vol. 5, pages 32-38. March, 1921.

——— "The Use of Intelligence Tests." *Educational Review*, Vol. 62, No. 2. September, 1921.

Colvin, Stephen S., and MacPhail, Andrew H. "Value of Psychological Tests at Brown University." *School and Society*, Vol. 16, pages 113-122. July 29, 1922.

DAGNEY, S. "Intelligence Tests and Collegiate Selection." *School and Society*, Vol. 15, pages 593-595. May 27, 1922.

Dean of Columbia College and Director of Columbia University Admissions. Annual Reports, 1918-1922, inclusive. (Address Secretary of Columbia University, New York City.)

——— "Intelligence Tests and the Harvard School of Business Administration." *School and Society*, Vol. 15, pages 163-164. February, 1922.

HOWE, CHARLES S. "Our Overcrowded Colleges." *School and Society*, Vol. 15, pages 55-60. January 14, 1922.

JOHNSTON, J. B. "Tests for Ability before College Entrance" *School and Society*, Vol. 15, pages 345-353. April, 1922 ("The University ought to advise incapable youth against waste of time and money in attempting what is beyond their powers.")

KELLEY and TERMAN. "Dr. Ruml's Criticism of Mental Test Methods." *Journal of Philosophy*, Vol. 18, pages 459-465. 1921.

KIRKPATRICK, E. A. "Intelligence Tests in Massachusetts Normal Schools." *School and Society*, Vol. 15, pages 55-60. January 14, 1922.

LAWS and BOWIE. "Intelligence Tests in Examinations for Junior Scholarships." *Journal of Experimental Pedagogy*, Vol. 6, pages 155-169. 1921.

LEMAN, H. B. "Forecasting Failures in College Classes." *School Review*, Vol. 30, pages 382-387. May, 1922. (Shows time saved for students by proper initial placement through use of intelligence and content examinations which have been experimentally related to local standards.)

Lists of intelligence and educational tests available for use of colleges, normal schools, and high schools. (Address World Book Company, Yonkers-on-Hudson, New York.)

MCCALL, WILLIAM ANDERSON. *How to Measure in Education.* The Macmillan Company, New York. 1922.

MOORE, H. T. "Three Types of Psychological Rating in Use with Freshmen at Dartmouth." *School and Society*, Vol. 13, pages 418-420. 1921.

MYERS, G. C. "Teachers vs. Mental Tests as Prophets of School Progress." *School and Society*, pages 300-303. September 9, 1922.

National Society for the Study of Education. 21st Year Book: *Intelligence Tests and Their Use.* Part I: The Nature, History, and General Principles of Intelligence Testing.

Thorndike, E. L. "Measurement in Education."

Colvin, S. S. "Principles Underlying the Construction and Use of Intelligence Tests."

Rugg, Harold O. "Statistical Methods Applied to Educational Testing."

Whipple, G. M. "An Annotated List of Group Intelligence Tests."

Part II: The Administrative Use of Intelligence Tests.

Holmes, H. W. "Intelligence Tests and Individual Progress in School Work."

Layton, W. K. "The Group Intelligence Testing Program of the Detroit Public Schools."

Davis, Helen. "The Use of Intelligence Tests in the Classification of Pupils in the Public Schools of Jackson, Michigan."

Rogers, Agnes L. "Measurement of the Abilities and Achievements of Children in the Lower Primary Grades."

Pintner, Rudolph. "The Significance of Intelligence Testing in the Elementary School."

Trabue, M. R. "The Use of Intelligence Tests in Junior High Schools."

Miller, W. S. "The Administrative Use of Intelligence Tests in the High School."

Gambrill, Bessie L. "Some Administrative Uses of Intelligence Tests in the Normal School."

Rogers, Agnes L. "The Use of Psychological Tests in the Administration of Colleges of Liberal Arts for Women."

Whipple, G. M. "Intelligence Tests in Colleges and Universities."

National Society for the Study of Education. 17th Year Book. Part II: History and Theory and Early Practice in Mental and Educational Testing. Public School Publishing Company, Bloomington, Illinois.

Peterson and Kuderna. "Army Alpha in Normal Schools." *School and Society*, Vol. 13, pages 476-480. 1921.

Powers, S. R. "Intelligence as a Factor in the Election of High School Subjects." *School Review*, Vol. 30, pages 452-455. June, 1922.

Terman, L. M. "The Psychological Determinist, or Democracy and the IQ." *Journal of Educational Research*, Vol. 6, No. 1, pages 57-62. June, 1922.

Rogers, Agnes L. "Intelligence Tests and Educational Progress." *Educational Review*, 1921.

——— "Use of Intelligence Tests in the Administration of Women's Colleges of Liberal Arts." 21st Year Book of the National Society for the Study of Education. Public School Publishing Company, Bloomington, Illinois.

Rogers, David Camp. "Intelligence Examinations and College Entrance." *The Smith Alumnae Quarterly*, Vol. 13, No. 1. November, 1921.

Rugg, H. O. *Statistical Methods Applied to Education.* Houghton Mifflin Company, Boston. 1917.

Scott, W. D. "Intelligence Tests for Prospective Freshmen." *School and Society*, Vol. 15, pages 382-388. April 8, 1922.

Stone, C. L. "The Significance of Alpha in Colleges." *Journal of Educational Psychology*, Vol. 13, page 218. May, 1922.

Terman, L. M. "Intelligence Tests in Colleges and Universities." *School and Society*, Vol. 13, pages 481-494. 1921.

Thorndike, E. L. *Introduction to the Theory of Mental and Social Measurements.* Teachers College, Columbia University. 1913.

TRABUE and STOCKBRIDGE. *Measure Your Mind.* Doubleday, Page & Co., Garden City, New York. 1921.

WHIPPLE, GUY M. "A Summary of the Work Done in the Testing of College Students." 21st Year Book of the National Society for the Study of Education, Part II, Chapter X, pages 253-270.

WHITECHURCH, A. R. "Psychological Norms among University Freshmen." *Journal of Applied Psychology,* Vol. 5, pages 318-339. 1921.

WILLIAMS, ROBERT DAY. "The Thorndike Intelligence Examinations of Pomona College Freshmen." *Pomona College Quarterly Magazine,* Vol. 10, No. 3. March, 1922.

WILSON, W. R. "Mental Tests and College Teaching." *School and Society,* Vol. 15, pages 629-635. June 10, 1922.

INDEX

Ability and achievement, importance of distinction between, 153
Academic work, as the chief purpose of the college, 110–112
Acquaintance, influence of personal, in marking examinations, 191–192
Administration of Thorndike Intelligence Examination, 302–305
Admission, proposed combination criterion for, 86–88
Algebra, College Entrance Board Examinations in, 124–131
Analysis of Thorndike Examination, 92–106
Army Alpha test, 23; distinction between Thorndike Examination and, 29–30; criticisms of, 159–160
Athletic competitions, as factor in college success, 108–109

Basis of college marks, the, 114–116
Brown University, Thorndike scores of Freshmen at (1919), 41
Butler, President, address by, on "The Great Tradition," 110–111

Cattell, J. McK., early testing work of, 2
Chadwick, Harold K., tables by, 70–72; study of "secondary checks" by, 93–98
Character, an element in determining fitness for college work, 17
Cheating, prevention of, in administration of Thorndike Examination, 54–57
Civil Engineering, trial of New Type examination in, 271–273
College marks, characteristics of, 112–113; basis of, 114–116; means of deriving, 116–117; derived from written examinations, 118–119; purpose of, 119–121
College success, discussion of, 107–140; effect of New Examination method on index of, 207–209. *See also* Success
Columbia College, use of intelligence tests in, 13–21; reliability of grades in, 131–134
Commuting, effect of, on college standing, 62, 79, 106
Comprehensive examination, use of, at Columbia, 18
Content examining, experiment with new method of, 177–214
Cost of giving Thorndike Examination, 57–58

Economics, applicability of New Type examination to, 250–256
Economy of New Examination method, 198–199
Educational measurement, principles of, 141–176
English, applicability of New Type examination to, 266–271

Failure, prediction of success and, in college, 63–68. *See* College marks
Fore-exercise, a feature of Thorndike Examination, 27–29

Geometry, College Entrance Board Examinations in, 124–131; New Type objective test for achievement in, 277
Government, experience with New Type examination in, 226–239
"Great Tradition, The," meaning of, 110–111
Greek Art, applicability of New Type examination to, 260–262

Hawkes, Herbert E., quoted on New Method of Admission at Columbia, 20–21
Health, an element in determining fitness for college work, 17
High school marks, poverty of, in predictive value, 81–86
History, applicability of New Type examination to, 262–265

Instructors, opinions of, on New Examination method, 201–205
Intelligence, an element in determining fitness for college work, 17
Intelligence scores, correlation of, with college success, 69–74
Intelligence tests, official status of, in Columbia College, 13–21. *See* Thorndike College Entrance Intelligence Examination

Jones, A. L., quoted on administration of intelligence examination at Columbia, 52

Kelley, T. L., quoted on reliability coefficients, 45, 47

Measurement, principles of educational, 141–176
Morphology, absence of relationship between college success and, 90–91

Naccarati, Santê, paper on "The Morphologic Aspect of Intelligence" by, 90–91
New Type Content Examination, experiment with, 177–214; experience of Department of Physics at Columbia with, 215–225; experience of Department of Government with, 226–239; experience of Department of Zoölogy with, 240–249; use of, in Economics, 250–256; in Philosophy, 256–260; in Greek Art, 260–262; in History, 262–265; in English, 266–271; in Civil Engineering, 271–273; future use of, 274–291

Objectivity, as an essential of valid scale for educational measurement, 141–145

Personal acquaintance, influence of, in marking examinations, 191–192
Phi Beta Kappa, use of pure quality index by, 75–76
Philosophy, applicability of New Type examination to, 256–260
Physical competitions, as factor in college success, 108–109
Physics, experience with New Type examination in, 215–225
Pomona College, Thorndike scores of Freshmen at, 41
Predictive values. *See* Success
Preparation, an element in determining fitness for college work, 17

Quality and quantity indexes, uses and comparative significance of, in calculating scholarship score, 75–80

Records, methods of keeping, 58–62, 279–280
Regents' examinations, predictive values of, as to college success or failure, 80–86; good criteria for admission to college, 82
Reliability coefficient, Thorndike's, 45

Scale for measurement of mental and social facts, essentials of a valid, 145
Scholarship score, defined, 74–75
School marks, inaccuracy of, 121–123

Secondary checks, deriving of, 92–93
Self-correlation of a test, 45
Stanford University, distributions of Thorndike Examination total scores for students at, 36–37
Starch, Daniel, *Educational Psychology* by, cited, 121
Students' Army Training Corps, use of Thorndike Test with, 15–16
Success, prediction of, in college, 63–68; correlation of intelligence scores with, 69–74; comparison of predictive values of new and old plans of admission, 80–86; absence of relationship between morphology and, 90–91. *See also* College success

Terman, L. M., data supplied by, 36–37
Thorndike, E. L., quoted on principles of educational measurement, 141; *Mental and Social Measurements* by, cited and quoted, 141 ff.
Thorndike College Entrance Intelligence Examination, 16; description of, 22 ff.; origin of, 22–25; characteristics of, 25–32; distributions of total scores, 32–42; stability of the mean and S. D. of, 42–44; reliability of, 45–51; method of conducting, 51–57; cost of giving, 57–58; methods of keeping records, 58–62; validity of, 63–91; analysis of, 92–106; sample questions of, 293–302; instructions for giving, 302–305; instructions for scoring, 305–307; data on separate parts of, 309–319; suggested improvements in, 320–330

Validity of Thorndike Examination, 63–91

Yerkes Committee, ten-test Army Examination *a* devised by, 23

Zero point, reference to a defined, as an essential of a valid scale for educational measurement, 145–150
Zoölogy, experience with New Type examination in, 240–249

Introduction to the Use of Standard Tests

By SIDNEY L. PRESSEY
and LUELLA C. PRESSEY
Ohio State University

A SIMPLE and practical manual making clear the fundamentals of effective testing. It is written in a non-technical, conversational style throughout and is well adapted to the needs of the busy teacher, supervisor, or superintendent beginning a testing program.

The first part, *How to Use Tests,* deals with the nature of tests, problems to be dealt with, methods of handling results, and mistakes to be avoided.

The second part, *Tests in School Subjects,* is a brief discussion of certain of the best tests in various school subjects and the problems of measurement in special subjects.

In the third part, *Tests of General Ability,* tests of mental ability, both individual and group, are discussed, with emphasis upon the practical uses of the results.

The fourth part, *Important General Principles,* points out many facts that the tester should know, including methods of making tests and the value of a systematic testing program.

A *Glossary* containing 87 technical terms and an *Index* are included.

Cloth. vi + 263 pages. Price $1.80

WORLD BOOK COMPANY
YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Especially designed for use in high schools
(Also usable as low as Grade 6 and as high as first year in college)

TERMAN GROUP TEST *of* MENTAL ABILITY

By LEWIS M. TERMAN

Professor of Educational Psychology, Stanford University; joint author of the National Intelligence Tests and of the army mental tests; author of the Stanford Revision of the Binet-Simon Scale, and of a number of books on the measurement of intelligence

This test is unique in many respects. Each of its 886 items was measured against a composite outside criterion. A try-out resulted in a reduction to 370 items, each helping to differentiate bright pupils from dull ones. The items retained are more highly selected than will be found in any other group mental test.

The Terman Test is an eleven-page booklet. The pupil does no writing. The backs of the Scoring Keys contain the scoring rules. Only 30 to 35 minutes will be required to test a group with it. The procedure has been so simplified that it can be mastered by any teacher in a few minutes. The size of the booklets makes their use without desks easy.

Examination: Form A. Price per package of 25 booklets, including Scoring Key and Manual of Directions, $1.35 net.

Examination: Form B. Price per package of 25 booklets, including Scoring Key and Manual of Directions, $1.35 net.

Specimen Set. Price 15 cents postpaid.

WORLD BOOK COMPANY

YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

21329

Date Due

DEC 1 1956

JAN 23 1957

MAY 6 1959

JUN 3 1959

JA 29'66

BF431
.W64
1923

ST. OLAF COLLEGE
BF4 .64 1
Wood
3 0111 00104 9633